Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer

Z-Image Team, Alibaba Group

Abstract

The landscape of high-performance image generation models is currently dominated by proprietary systems, such as Nano Banana Pro [27] and Seedream 4.0 [64]. Leading open-source alternatives, including Qwen-Image [76], Hunyuan-Image-3.0 [8] and FLUX.2 [36], are characterized by massive parameter counts (20B to 80B), making them impractical for inference, and fine-tuning on consumer-grade hardware. To address this gap, we propose **Z-Image**, an efficient *6B-parameter* foundation generative model built upon a Scalable Single-Stream Diffusion Transformer (S3-DiT) architecture that *challenges the "scale-at-all-costs" paradigm*. By systematically optimizing the entire model lifecycle – from a curated data infrastructure to a streamlined training curriculum – we complete the full training workflow in *just 314K H800 GPU hours (approx. $630K)*. Our few-step distillation scheme with reward post-training further yields **Z-Image-Turbo**, offering both *sub-second inference latency* on an enterprise-grade H800 GPU and *compatibility with consumer-grade hardware (<16GB VRAM)*. Additionally, our omni-pre-training paradigm also enables efficient training of **Z-Image-Edit**, an editing model with impressive instruction-following capabilities. Both qualitative and quantitative experiments demonstrate that our model achieves performance comparable to or surpassing that of leading competitors across various dimensions. Most notably, Z-Image exhibits *exceptional capabilities in photorealistic image generation and bilingual text rendering*, delivering results that rival top-tier commercial models, thereby demonstrating that *state-of-the-art results are achievable with significantly reduced computational overhead*. We publicly release our code, weights, and online demo to foster the development of accessible, budget-friendly, yet state-of-the-art generative models.

⌂	**GitHub**	https://github.com/Tongyi-MAI/Z-Image
⟨⟩	**ModelScope Model**	https://modelscope.cn/models/Tongyi-MAI/Z-Image-Turbo
🤗	**HuggingFace Model**	https://huggingface.co/Tongyi-MAI/Z-Image-Turbo
⟨⟩	**ModelScope Demo**	Online Demo (ModelScope)
🤗	**HuggingFace Demo**	Online Demo (HuggingFace)
🖼	**Image Gallery**	Online Gallery Offline Gallery

Figure 1 | Showcases of Z-Image-Turbo in photo-realistic image generation.

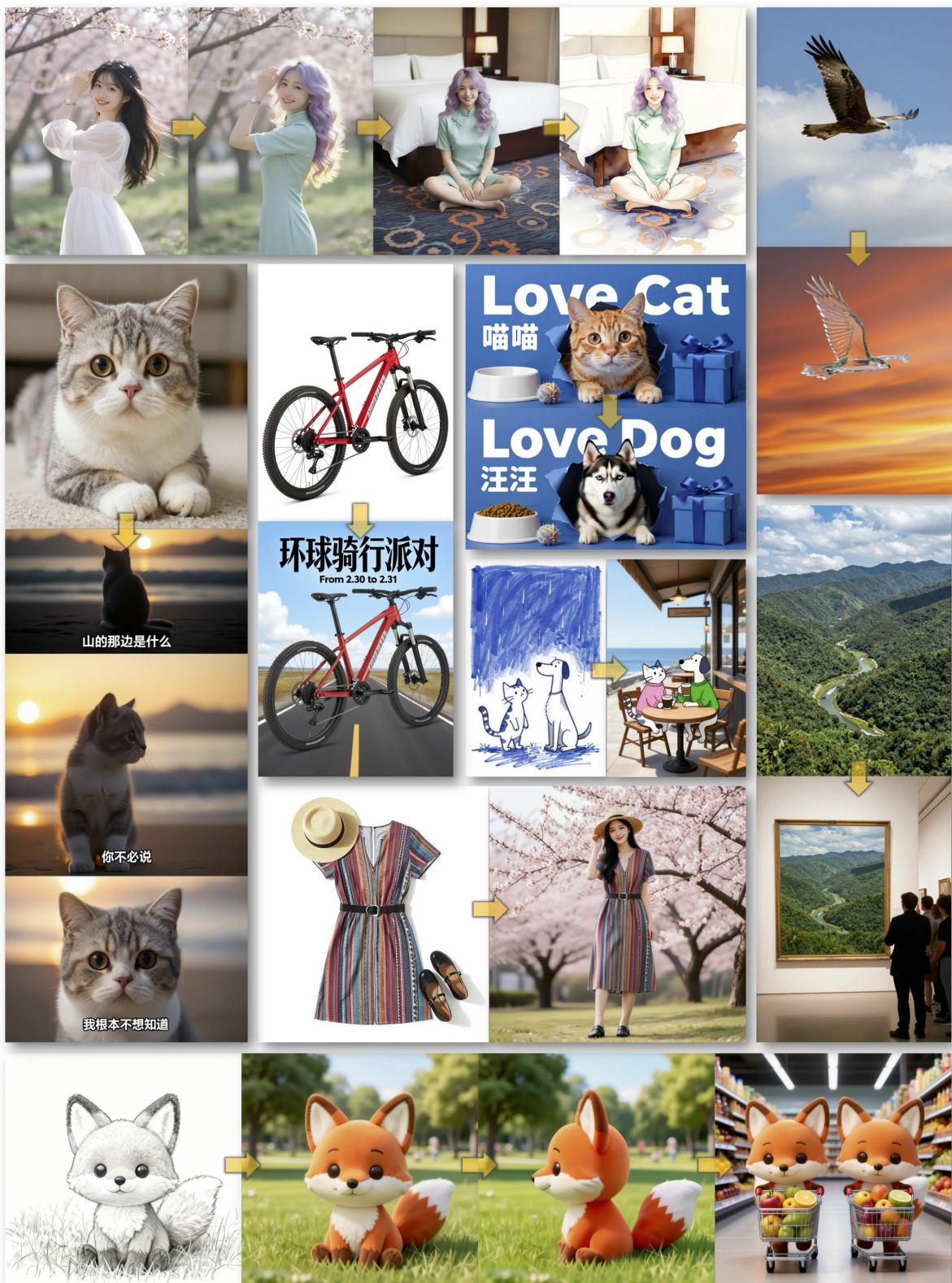Figure 2 | Showcases of Z-Image-Turbo in bilingual text-rendering.

Figure 3 | Showcases of Z-Image-Edit in various image-to-image tasks. Each arrow represents an edit from the input to output images.

|	
Input prompt	Under a sudden iPhone flash, she stands just where the surf and sand meet, the midi-skirt and fluid blouse catching the gentle seaside breeze, woven folds reflecting faint glimmers of distant city lights and moon's shy glow. Her bare feet nestle softly in the warm, textured sand, illuminated subtly by distant streetlamps that drift behind her, palms and skyscrapers whispering quiet stories to the night. Damp tendrils of wavy hair move gently, framing a serene, turned face softened by reflective air, while the expansive darkness spills across two-thirds of the frame in velvety repose. A delicate blur trails her skirt's hem, contrasting with the crisp focus of her poised stance, embodying timeless grace and effortless charm—flash-lit elegant beach snapshot, captured on iPhone.
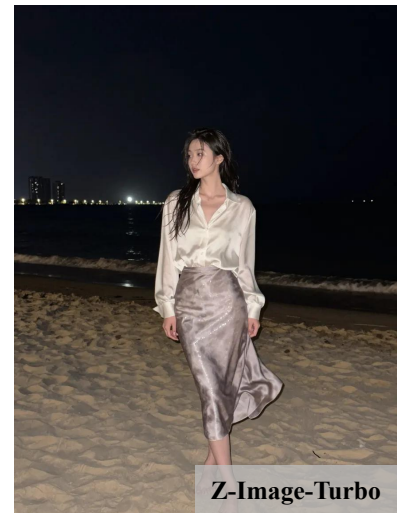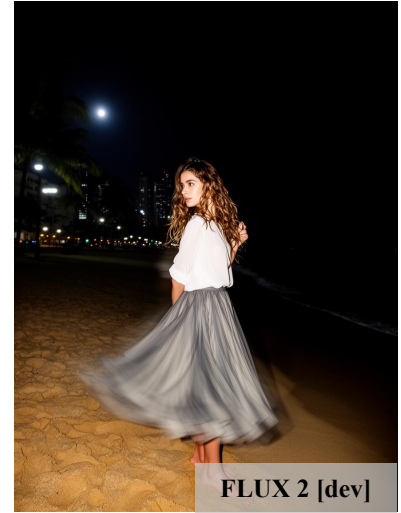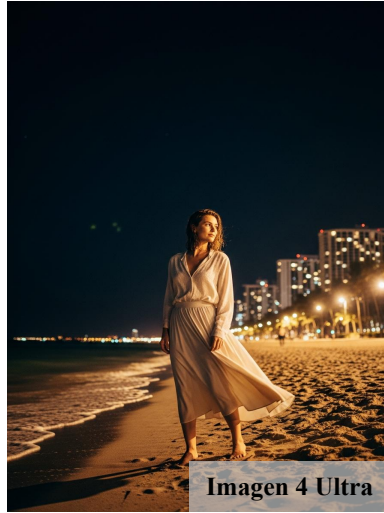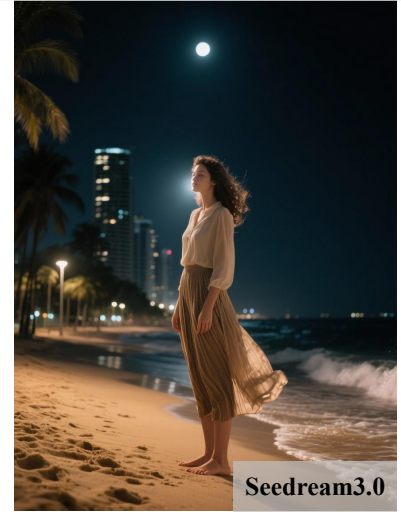
Figure 4 | Showcases of comparison between Z-Image-Turbo and currently state-of-the-art models [58, 76, 8, 27, 36, 64, 21, 26]. Z-Image-Turbo shows conspicuous photo-realistic generation capacity.

5

Contents

1. Introduction

The field of text-to-image (T2I) generation has witnessed remarkable advancements in recent years, evolving from generating rudimentary textures to producing photorealistic imagery with complex semantic adherence [57, 18, 35, 76, 64, 8, 4]. However, as the capabilities of these models have scaled, their development and accessibility face significant barriers. The current landscape is increasingly characterized by two divergent trends: on one side, state-of-the-art commercial closed-source models – such as Nano Banana Pro [27] and Seedream 4.0 [64] – remain enclosed within "black boxes", offering high performance but limited transparency or reproducibility. On the other side, open-source models, while fostering democratization, often resort to massive parameter scaling – approaching tens of billions of parameters (*e.g.*, Qwen-Image [76] (20B), FLUX.2 [36] (32B) and Hunyuan-Image-3.0 [8] (80B) – imposing prohibitive computational costs for both training and inference. In this context, distilling synthetic data from proprietary models has emerged as an appealing shortcut to train high-performing models at lower cost, becoming a prevalent approach for resource-constrained academic research [13, 20]. However, this strategy risks creating a closed feedback loop that may lead to error accumulation and data homogenization, potentially hindering the emergence of novel visual capabilities beyond those already present in the teacher models.

In this work, we present **Z-Image**, a powerful diffusion transformer model that challenges both the "scale-at-all-costs" paradigm and the reliance on synthetic data distillation. We demonstrate that neither approach is necessary to develop a top-tier image generation model. Instead, we introduce the first comprehensive end-to-end solution that systematically optimizes every stage of the model lifecycle – from data curation and architecture design to training strategies and inference acceleration – enabling efficient, low-cost development on **purely real-world data without distilling results from other models**.

Most notably, this methodological efficiency allows us to complete the entire training workflow with remarkably low computational overhead. As detailed in Table 1, the complete training pipeline for Z-Image requires only **314K H800 GPU hours**, translating to approximately **$628K** at current market rates (about $2 per GPU hour [38]). In a landscape where leading models often demand orders of magnitude more resources, this modest investment demonstrates that principled design can effectively rival brute-force scaling.

Table 1 | Training costs of Z-Image, assuming the rental price of H800 is about $2 per GPU hour. The rental price refers from [38].

Training Costs	Low-res. Pre-Training	Omni-Pre-Training	Post-Training	Total
in H800 GPU Hours	147.5K	142.5K	24K	314K
in USD	$295K	$285K	$48K	$628K

This breakthrough in cost-efficiency is underpinned by a systematic methodology built on four pillars:

- **Efficient Data Infrastructure:** In resource-constrained scenarios, an efficient data infrastructure is pivotal; it serves to maximize the rate of knowledge acquisition per unit of time – thereby accelerating training efficiency – while simultaneously establishing the upper bound of model capabilities. To achieve this, we introduce a comprehensive Data Infrastructure composed of four synergistic modules: a *Data Profiling Engine* for multi-dimensional feature extraction, a *Cross-modal Vector Engine* for semantic deduplication and targeted retrieval, a *World Knowledge Topological Graph* for structured concept organization, and an *Active Curation Engine* for closed-loop refinement. By granularly profiling data attributes and orchestrating the training distribution, we ensure that the "right data" is aligned with the "right stage" of model development. This infrastructure maximizes the utility of real-world data streams, effectively eliminating computational waste arising from redundant or low-quality samples.
- **Efficient Architecture:** Inspired by the remarkable scalability of decoder-only architectures in large language models [6], we propose a **Scalable Single-Stream Multi-Modal Diffusion Transformer (S3-DiT)**. Unlike dual-stream architectures that process text and image modalities in isolation, our design facilitates dense cross-modal interaction at every layer. This high parameter efficiency enables Z-Image to achieve superior performance within a compact 6B parameter size, significantly

lowering the hardware requirements for both training and deployment. The compact model size is also made possible in part by our use of a prompt enhancer (PE) to augment the model's complex world knowledge comprehension and prompt understanding capabilities, further mitigating the limitations of the modest parameter count. Furthermore, this early-fusion transformer architecture ensures superior versatility by treating tokens from different modalities uniformly – including text tokens, image VAE tokens, and image semantic tokens – enabling seamless handling of diverse tasks such as text-to-image generation and image-to-image editing within a unified framework.

- **Efficient Training Strategy:** We design a progressive training curriculum composed of three strategic phases: (1) **Low-resolution Pre-training**, which bootstraps the model to acquire foundational visual-semantic alignment and synthesis knowledge at a fixed 256^2 resolution. (2) **Omni-pre-training**, a unified multi-task stage that consolidates arbitrary-resolution generation, text-to-image synthesis, and image-to-image manipulation. By amortizing the heavy pre-training budget across these diverse capabilities, we eliminate the need for separate, resource-intensive stages. (3) **PE-aware Supervised Fine-tuning**, a joint optimization paradigm where Z-Image is fine-tuned using PE-enhanced captions. This ensures seamless synergy between the Prompt Enhancement module and the diffusion backbone without incurring additional LLM training costs, thereby maximizing the overall development efficiency of the Z-Image system.

- **Efficient Inference:** We present **Z-Image-Turbo**, which delivers exceptional aesthetic alignment and high-fidelity visual quality in only 8 Number of Function Evaluations (NFEs). This performance is unlocked by the synergy of two key innovations: *Decoupled DMD* [46], which explicitly disentangles the quality-enhancing and training-stabilizing roles of the distillation process, and *DMDR* [32], which integrates Reinforcement Learning by employing the distribution matching term as an intrinsic regularizer. Together, these techniques enable highly efficient generation without the typical trade-off between speed and quality.

Building upon this robust foundation and efficient workflow, we have successfully derived two specialized variants that address distinct application needs. First, our few-shot distillation scheme with reinforcement learning yields **Z-Image-Turbo**, an accelerated model that achieves exceptional aesthetic alignment in just 8 NFEs. It offers **sub-second inference**[1] **latency on enterprise GPUs** and **fits within the memory constraints of consumer-grade hardware (<16GB VRAM)**. Second, leveraging the multi-task nature of our omni-pre-training, we introduce **Z-Image-Edit**, a model specialized for precise instruction-following image editing.

Extensive qualitative and quantitative experiments demonstrate the superiority of the Z-Image family. As illustrated in Figure 1 and Figure 2, Z-Image delivers strong capabilities of photorealistic generation and exceptional bilingual (Chinese/English) text rendering, matching the visual fidelity of much larger models. Figure 3 showcases the capabilities of Z-Image-Edit, highlighting its precise adherence to editing instructions. Furthermore, qualitative comparisons in Figure 4 and Section 5.3 reveal that our model rivals top-tier commercial systems, proving that **state-of-the-art results are achievable with significantly reduced computational overhead**. We publicly release our code, weights, and online demo to foster the development of accessible, budget-friendly generative models.

2. Data Infrastructure

While the remarkable capabilities of state-of-the-art text-to-image models are underpinned by large-scale training data, achieving optimal performance under constrained computational resources necessitates a paradigm shift from data quantity to data efficiency. Simply scaling the dataset size often leads to diminishing returns; instead, an efficient training pipeline requires a data infrastructure that maximizes the information gain per computing unit. To this end, an ideal data system must be strictly curated to be *conceptually broad* yet *non-redundant*, exhibit *robust multilingual text-image alignment*, and crucially, be *structured for dynamic curriculum learning*, ensuring that the data composition evolves to match the model's training stages. To realize this, we have designed and implemented an integrated Efficient Data Infrastructure. Far from a static repository, this system operates as a dynamic engine architected to maximize the rate of knowledge acquisition within a fixed training budget. As the cornerstone of our pipeline, this infrastructure is composed of four core, synergistic modules:

---

[1]FlashAttention-3 [65] and `torch.compile` [1] is necessary for achieving sub-second inference latency.

1. **Data Profiling Engine:** This module serves as the quantitative foundation for our data strategy. It extracts and computes a rich set of multi-dimensional features from raw data, spanning low-level physical attributes (*e.g.*image metadata, clarity metrics) to high-level semantic properties (*e.g.*, anomaly detection, textual description). These computed profiles are not merely for basic filtering; they are the essential signals used to quantify data complexity and quality, enabling the programmatic construction of curricula for our dynamic learning stages.

2. **Cross-modal Vector Engine:** Built on billions of embeddings, this module is the engine for ensuring efficiency and diversity. It directly supports our goal of a *non-redundant* dataset through large-scale semantic deduplication. Furthermore, its cross-modal search capabilities are critical for diagnosing and remediating model failures. This allows us to pinpoint and prune data responsible for specific failure cases and strategically sample to fill conceptual gaps.

3. **World Knowledge Topological Graph:** This structured knowledge graph provides the semantic backbone for the entire infrastructure. It directly underpins our goal of *conceptual breadth* by organizing knowledge hierarchically. Crucially, this topology functions as a semantic compass for data curation. It allows us to identify and fill conceptual voids in our dataset by traversing the graph to find underrepresented entities. Furthermore, it provides the structured framework needed to precisely rebalance the data distribution across different concepts during training, ensuring a more efficient and comprehensive learning process.

4. **Active Curation Engine:** This module operationalizes our infrastructure into a truly dynamic, self-improving system. It serves two primary, synergistic functions. First, it acts as a frontier exploration engine, employing automatic sampling to identify concepts on which the model performs poorly or lacks knowledge ("hard cases"). Second, it drives a closed-loop data annotation pipeline. This ensures that every iteration not only expands *conceptual breadth* of the dataset with high-value knowledge but also continuously refines the data quality, maximizing the learning efficiency of the entire training process.

Collectively, these components forge a robust data infrastructure that not only fuels the training of text-to-image models but also establishes a versatile infrastructure for broader multimodal model training. Leveraging this system, we successfully facilitate the training of various critical components, including captioners, reward models, and our image editing model (*i.e.*, Z-Image-Edit). In particular, we construct a dedicated data pipeline specifically for Z-Image-Edit upon this infrastructure, the details of which are elaborated in Section 2.5.

2.1. Data Profiling Engine

The Data Profiling Engine is engineered to systematically process a massive, uncurated data pool, comprising large-scale internal copyrighted collections. It computes a comprehensive suite of multi-dimensional features for each image-text pair, enabling principled data curation. Recognizing that different data sources exhibit unique biases, our engine supports source-specific heuristics and sampling strategies to ensure a balanced and high-quality training corpus. The profiling process is structured across several key dimensions:

Image Metadata. We begin by caching fundamental properties for each image. This includes elementary metadata like resolution (width and height) and file size, which facilitate efficient filtering based on resolution and aspect ratio. Simultaneously, we compute a perceptual hash (pHash) from the image's byte stream. This hash acts as a compact visual fingerprint, enabling rapid and effective low-level deduplication to remove identical or near-identical images. Together, these pre-computed attributes form the first layer of data selection.

Technical Quality Assessment. The technical quality of an image is a critical determinant of model performance. Our engine employs a multi-faceted approach to quantify and filter out low-quality assets:

- **Compression Artifacts:** To identify over-compressed images, we calculate the ratio of the ideal uncompressed file size (derived from resolution and bit depth) to the actual file size. A low ratio indicates potential quality degradation due to excessive compression.
- **Visual Degradations:** We utilize an in-house trained quality assessment model to score images

on a range of degradation factors, including color cast, blurriness, perceptible watermarks, and excessive noise.

- **Information Entropy:** To maximize the density of meaningful content seen during training, we filter out low-entropy images. This is achieved through two complementary methods: (1) analyzing the variance of border pixels to detect images with large, uniform-color backgrounds or frames, and (2) performing a transient JPEG re-encoding and using the resulting bytes-per-pixel (BPP) as a proxy for image complexity.

Semantic and Aesthetic Content. Beyond technical quality, we profile the high-level semantic and aesthetic properties of images:

- **Aesthetic Quality:** We leverage an aesthetics scoring model, trained on labels from professional annotators, to quantify the visual appeal of each image.
- **AIGC Content Detection:** Following the findings of Imagen 3 [3], we trained a dedicated classifier to detect and filter out AI-generated content. This step is crucial for preventing degradation in the model's output quality and physical realism.
- **High-Level Semantic Tagging:** We have trained a specialized Vision-Language Model (VLM) to generate rich semantic tags. These tags include general object categories, human-centric attributes (*e.g.*number of people), and culturally specific concepts, with a particular focus on elements relevant to Chinese culture. The same model also performs safety assessment by assigning Not-Safe-for-Work (NSFW) scores, allowing for the unified filtering of both semantically irrelevant and inappropriate content.

Cross-Modal Consistency and Captioning. The alignment between an image and its textual description is paramount.

- **Text-Image Correlation:** We use CN-CLIP [86] to compute the alignment score between an image and its associated alt caption. Pairs with low correlation scores are discarded to ensure the relevance of textual supervision.
- **Multi-Level Captioning:** For all images selected for pre-training, we generate a structured set of captions, including concise tags, short phrases, and detailed long-form descriptions. Notably, diverging from prior works [21, 64, 76] that use separate modules for Optical Character Recognition (OCR) and watermark detection, our approach leverages the powerful inherent capabilities of our VLM. We explicitly prompt the VLM to describe any visible text or watermarks within the image, seamlessly integrating this information into the final caption. This unified strategy not only streamlines the data processing pipeline but also enriches the textual descriptions with critical visual details, as further elaborated in Section 3.

2.2. Cross-modal Vector Engine

We enhance the de-duplication method proposed in Stable Diffusion 3 [18], reformulating it as a scalable, graph-based community detection task. Addressing the severe scalability bottleneck of the original *range_search* function, we substitute it with a highly efficient k-nearest neighbor (k-NN) *search* function. We construct a proximity graph from the k-NN distances and subsequently apply the community detection algorithm [68]. This methodology closely approximates the original algorithm's output for a sufficiently large k while drastically reducing time complexity. Our fully GPU-accelerated [60] pipeline achieves a processing rate of approximately 8 hours per 1 billion items on 8 H800s, encompassing index construction and 100-NN querying. This approach not only ensures a non-redundant dataset by identifying dense clusters for effective de-duplication but also extracts semantic structures via modularity levels, facilitating fine-grained data balancing.

Furthermore, we constructed an efficient retrieval pipeline leveraging multimodal features [86] combined with a state-of-the-art index algorithm [54]. This system's cross-modal search capabilities are critical for both data curation and active model remediation. Beyond identifying distributional voids for strategically sampling to fill conceptual gaps – thereby enabling targeted augmentation for a balanced pre-training distribution – this engine is instrumental in diagnosing model failures. By querying the

system with failure cases (*e.g.*, problematic generated images or text prompts), we can pinpoint and prune the underlying data clusters responsible for the erroneous behavior. This iterative refinement process, targeting both data gaps and model failures, ensures dataset robustness and is pivotal for sourcing high-quality candidates for complex downstream tasks.

2.3. World Knowledge Topological Graph

The construction of our knowledge graph follows a three-stage process. Initially, we build a comprehensive but redundant knowledge graph from all Wikipedia entities and their hyperlink structures. To refine this graph, we employ a two-pronged pruning strategy: first, centrality-based filtering removes nodes with exceptionally low PageRank [56] scores, which represent isolated or seldom-referenced concepts; second, visual generatability filtering uses a VLM to discard abstract or ambiguous concepts that cannot be coherently visualized. Subsequently, to address the limited conceptual coverage of the pruned graph, we augment it by leveraging a large-scale internal dataset of captioned images. We extract tags and corresponding text embeddings from all available captions. Inspired by [71], we then perform an automatic hierarchical strategy on these embeddings. Each parent node is named by using a VLM to summarize its child nodes. This not only supplements the graph with new concept nodes but also organizes them into a structured taxonomic tree, significantly enhancing the structural integrity of the graph. In the final stage, we perform weight assignment and dynamic expansion to align the graph with practical applications. This involves manually curating and up-weighting high-frequency concepts from user prompts, and proactively integrating novel, trending concepts not yet present in our data pool to maintain the relevance and timeliness of the graph.

In application, this graph underpins our semantic-level balanced sampling strategy. We map the tags within each training caption to their corresponding nodes in the knowledge graph. By considering both the BM25 [62] score of a tag and its hierarchical relationships (*i.e.*, parent-child links) within the graph, we compute a semantic-level sampling weight for each data point. This weight then guides our data engine to perform principled, staged sampling from the data pool, enabling fine-grained control over the training data distribution.
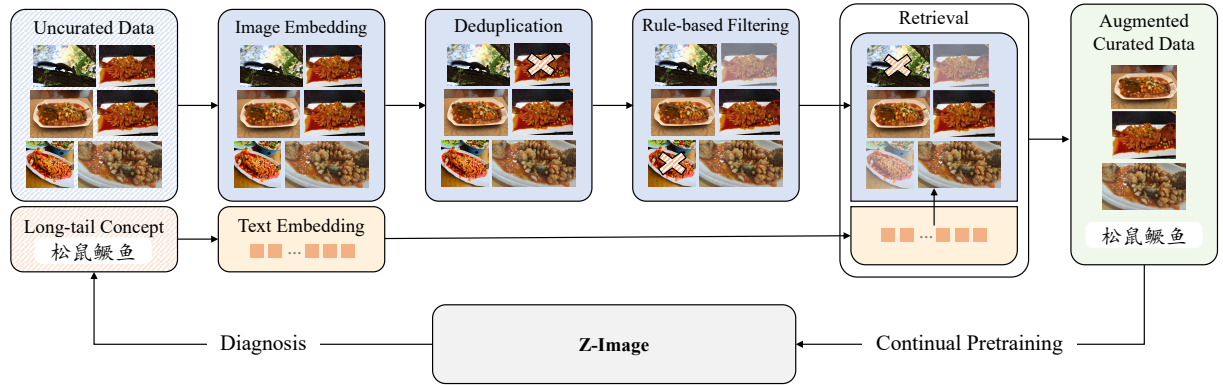
Figure 5 | Overview of the Active Curation Engine. The pipeline refines uncurated data through cross-modal embedding, deduplication, and rule-based filtering to construct a high-quality augmented dataset. A feedback mechanism leverages the Z-Image model to diagnose long-tail distribution deficiencies, dynamically guiding cross-modal retrieval to reinforce the data collection process. The "Squirrel Fish" (松鼠鳜鱼) case illustrates a classic long-tail challenge: it is actually the name of a Chinese cuisine but the model lacks the specific concept for this dish and may rely on compositional reasoning (combining "Squirrel" (松鼠) and "Fish" (鳜鱼)), leading to erroneous generations absent of domain-specific training data.

2.4. Active Curation Engine

To systematically elevate data quality and address long-tail distribution challenges, we deploy a comprehensive Active Curation Engine (Figure 5). This framework incorporates a filtering tool and Z-Image as
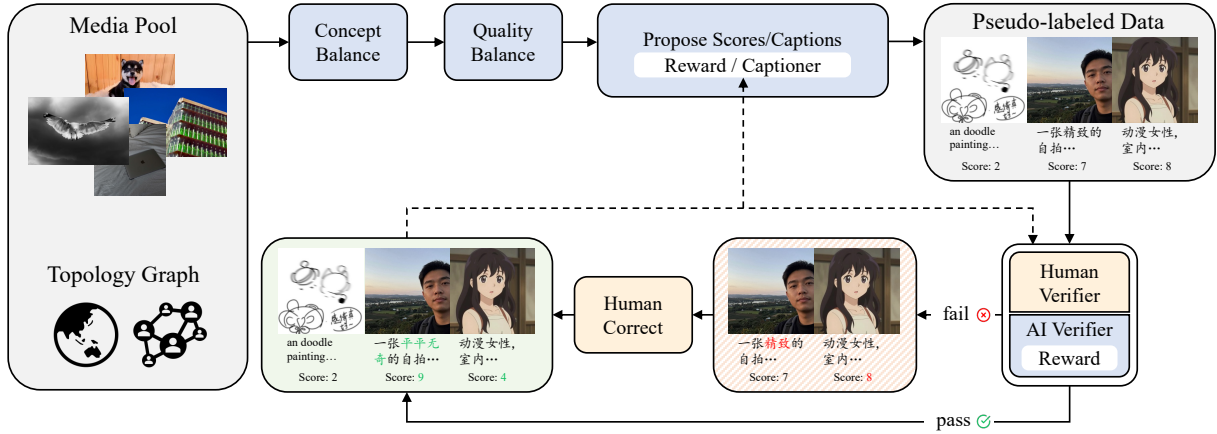
Figure 6 | Illustration of the Human-in-the-Loop Active Learning Cycle. Data sampled from the media pool undergoes concept and quality balancing before being assigned pseudo-labels . A dual-verifier system (Human and AI) filters these proposals: approved samples pass directly, while rejected cases trigger a manual correction phase . This feedback loop iteratively refines the annotations and updates the topology graph to ensure high-precision alignment.

a diagnostic generative prior. The pipeline begins by processing uncurated data through cross-modal embedding and deduplication, followed by rule-based filtering to eliminate low-quality samples.

To support the continuous evolution of Z-Image, we establish a human-in-the-loop active learning cycle (Figure 6) where the reward model and captioner are progressively optimized. In this pipeline, we first employ the topology graph (Section 2.3) and the initial reward model to curate a balanced subset from the unlabeled media pool. The current captioner and reward model then assign pseudo-labels to these samples. A hybrid verification mechanism – comprising both human and AI verifiers – verifies these proposals; rejected samples trigger a manual correction phase by human experts to refine captions or scores. This high-quality annotated data is then used to retrain the captioner and reward model, thereby creating a virtuous cycle of our whole data infrastructure enhancement.

2.5. Efficient Construction of Editing Pairs with Graphical Representation

(a) Graphical Representation   (b) Paired Image from Videos   (c) Rendering for Text Editing

Figure 7 | Data construction for image editing using different strategies: (a) arbitrarily permuting and combining different edited versions of the same input image where the green arrow represents the pair constructed by task-specfic expert models and the red arrow denotes the pair generated by combination and permutation, (b) collecting images with inherent relationship from video frames, and (c) controllable text rendering system for text editing.

Collecting editing pairs that exhibits precise instruction following is challenging, owing to the requirement of consistency maintaining and the diverse and complex nature of editing operations. Through scalable

and controllable strategies as shown in Figure 7, we construct a large-scale training corpus from diverse sources.

Mixed Editing with Expert Models. To guarantee broad task coverage, we begin by curating a diverse taxonomy of editing tasks, and then leverage task-specific expert models to synthesize high-quality training data for each category. To improve the training efficiency, we construct mixed-editing data, where multiple editing actions are integrated into one editing pair. Thus, the model can enhance its ability in multiple editing tasks from only a single composite pair, instead of relying on multiple ones.

Efficient Graphical Representation. For an input image, we synthesize multiple edited versions corresponding to different editing tasks, enabling us to further scale the training data at zero cost through arbitrary pairwise combination [42] (*e.g.*, $2\binom{N+1}{2}$ pairs are constructed from one input image and its N edited versions). Apart from scaling the quantity, this strategy 1) creates mixed-editing training data by combining two edited versions to enhance the training efficiency, and 2) yields inverse pairs to improve data quality, *i.e.*, transforming a real, undistorted input image to an output image.

Paired Images from Videos. Constructing image editing pairs from predefined tasks suffers from limited diversity. To overcome this issue, we leverage naturally grouped images collected from a large scale video frames in our media pool. These images, by sharing inherent relatedness (*e.g.*, common subjects, scenes, or styles), implicitly define complex editing relationships among themselves. Building on this, we refine the data by calculating the cosine similarity between image embeddings using CN-CLIP [86], allowing us to filter for pairs with high semantic relevance within each image group. The resulting dataset of video frame pairs offers three key advantages: 1) high task diversity, 2) inherent coupling of multiple edit types (*e.g.*, simultaneous changes in human pose and background), and 3) superior scalability.

Rendering for Text Editing. The acquisition of high-quality training data for text editing presents substantial challenges, where natural images suffer from the scarcity and imbalance of textual content, and text editing requires paired samples with precise operation annotations. To address these challenges, we develop a controllable text rendering system [76] that grants us precise control over not only the textual content but also its visual attributes, such as font, color, size, and position. This approach enables us to systematically generate a large-scale dataset of paired images, where the ground-truth editing instruction are known by the rendering operation, thereby directly overcoming the aforementioned data limitations.
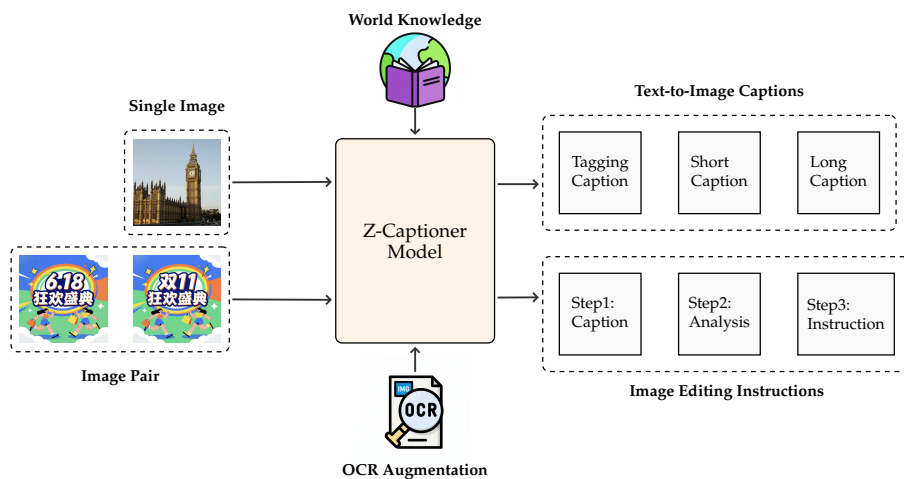
3. Image Captioner

Figure 8 | Pipeline for generating text-to-image captions and image editing instructions. OCR results (obtained through CoT) and world knowledge (from meta information) are explicitly included into the captions.

We build an all-in-one image captioner, Z-Captioner, by incorporating multiple types of image caption.

Tagging Caption:
1 girl, single subject focus, young woman, another person in the background, East Asian, black hair, bangs, white T-shirt, red apron, Hangzhou food festival, holding a plate, using chopsticks, Longjing shrimp, West Lake, Hangzhou, China, Leifeng Pagoda, EUROSTAR, magazine cover, outdoors, daytime, lakeside, water, tour boat, willow trees, displayed food, osmanthus cake, tea leaves, glass jars, orange triangular flags, white tents, photography, half-body shot, white border, 'Food', 'Culture', 'Design Travel', 'ND', '2025', 'THE MAGAZINE', 'F BIODI', 'FOOD', 'EUROSTAR', 'SPECIAL', 'TASTY HANGZHOU', '美食节攻略', 'HANGZOU FOOD FESTIVAL 2025', '骑行路线', '2025 杭州美食节'

Short Caption:
A promotional magazine cover themed around Hangzhou Cuisine, featuring a young woman at a West Lake food stall during the 2025 Hangzhou Food Festival, holding a plate and selling Longjing shrimp.

Long Caption:
A "Food ND" magazine cover, a special issue about "HANGZHOU FOOD FESTIVAL 2025."
The scene takes place on the shore of West Lake, one of Hangzhou's iconic landmarks, captured in a vibrant, sunny daytime setting. At the center of the image is a young East Asian woman with black hair, wearing a white T-shirt and a bright red apron. She is holding a white plate of the famous local dish Longjing Shrimp and lifting a piece with a pair of green chopsticks. The red apron is printed with white text reading "2025 杭州美食节". To her right is a food stall under a white tent, decorated with a string of orange triangular flags. On the table are various local delicacies: two jars of Longjing tea leaves, a wooden tray with square white cakes sprinkled with osmanthus, several jars of pickled vegetables, and an orange hanging tag with vertical black text reading "中雨香". In the background is the expansive view of West Lake, its surface shimmering in the sunlight, with several traditional-style boats gliding across the water. On the opposite shore rise lush green hills, and the famous Leifeng Pagoda stands clearly on the right-side hilltop...

Step1 Caption:
Source image: A high-quality close-up portrait of a domestic cat. It has white fur on its chest and around its mouth and nose, and striking green eyes that stare directly into the camera with a neutral expression. The lighting is soft and even, highlighting the details of its fur and whiskers...
Target image: An anthropomorphic cat is walking on a tropical sandy beach. The cat's head is exactly the same as in the original image, but its body is now dressed in a formal navy-blue pinstripe suit with a white shirt and a black bow tie. It stands on its hind legs, holding a glass of amber-colored liquid in its right paw...

Step2 : Analysis
Subject modification: The cat's body has been replaced with an upright, anthropomorphic body dressed in a navy-blue pinstripe suit, a white dress shirt, and a black bow tie.
Element addition: A glass containing an amber-colored liquid has been added to the cat's right paw.
Scene change: The entire background has been replaced...

Step3 : Instruction
Place this cat on a tropical beach. Replace its body with an upright, anthropomorphic one dressed in a navy-blue pinstripe suit, a white shirt, and a black bow tie. As it walks toward the camera, have it hold a glass of wine in its right paw.
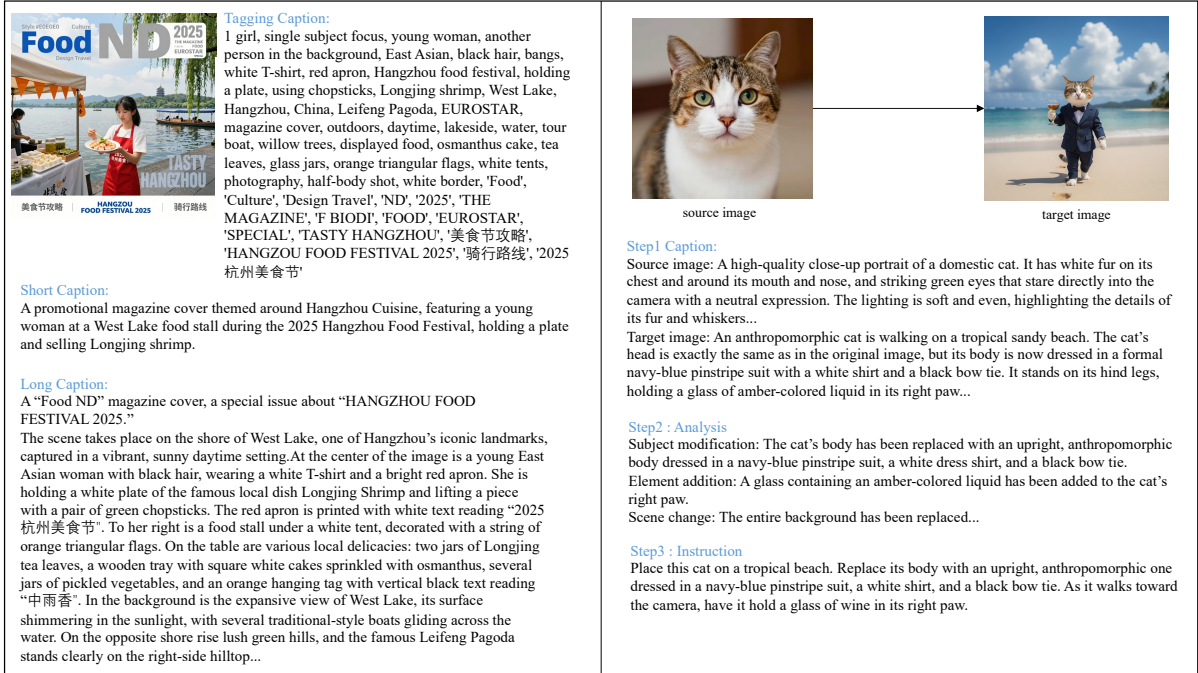
Figure 9 | Single image caption and difference caption examples. Left: for single image, we have captions of different types and lengths, and notably, OCR results (all the texts transcribed in their original languages) and world knowledge (explicitly and correctly recognizing the famous beauty spot, West Lake, Hangzhou, China, in this example) is included. Right: difference captions are composed step-by-step.

As revealed in previous works [49], different captioning tasks can benefit each other as they share the same goal of understanding and depicting images. Our model is designed not only to describe visual elements, but also to leverage extensive world knowledge to interpret the semantic context of the image. The integration of world knowledge is particularly critical for the downstream text-to-image synthesis task, as it enables the model to accurately render images involving specific named entities. Figure 8 shows our pipeline for generating text-to-image captions and image editing instructions.

3.1. Detailed Caption with OCR Information

First, we specially emphasize that according to our experiments, including explicit OCR information in image captions is inextricably bound with accurate text rendering in the generated images. Therefore, we employ a way that shares the same spirit as Chain-of-Thought (CoT) [73], by first explicitly recognizing all optical characters in the image and then generating a caption based on the OCR results. This effectively mitigates missing texts compared to directly generating a caption that encapsulates everything, especially for the cases where texts are very long/dense. In addition, we force the OCR results to remain in their original languages without any translation, avoiding them being falsely rendered in their translated languages.

3.2. Multi-Level Caption with World Knowledge

We design five different types of image captions in total, including long, medium and short captions, as well as tags and simulated user prompts. With the data infrastructure in Section 2, we include world knowledge in all five types of captions by performing image captioning conditioned on meta information. This significantly alleviates hallucinations when our captioner identifies and names specific entities such as public figures, famous landmarks, or known events.

To be specific, for relatively long captions, we include very dense information of the images, in order that the model could learn a mapping from the text to the image as accurate as possible. These captions contain full OCR results as mentioned above, along with subjects, objects, background, location information, et al.

We deliberately adopt a plain and objective linguistic style for our descriptions, strictly confining them to factual information observable in the image. By inhibiting subjective interpretations and imaginative associations, our purpose is to enhance data efficiency for the image generation task by eliminating non-essential information.

On the other hand, short captions, tags and simulated user prompts are designed for the model to adapt to real user prompts (which are usually short and unspecific) for better user experience. Notably, most of the simulated user instructions are incomplete prompts. They differ from short captions in that a short caption provides a relatively complete and comprehensive description of the entire image. In contrast, a short simulated prompt may mimic user behavior by focusing only on specific parts of interest to the user, while making no mention of the rest of the image.

3.3. Difference Caption for Image Editing

Difference caption is a concise editing instruction specifying the transformation from a source to a target image. To generate this, we employ a three-step CoT process that systematically breaks down the comparative task [100].

1. **Step1: Detailed Captioning.** We first generate a comprehensive, OCR-inclusive caption for both the source and target images respectively. This step provides a structured, detailed representation of each image's content.
2. **Step2: Difference Analysis.** The model then performs a comparative analysis, leveraging both the raw images and their generated captions, to tell all discrepancies from visual and textual perspectives.
3. **Step3: Instruction Synthesis.** Finally, the model generates a concise editing instruction based on the identified differences.

This step-by-step process helps the model create clear and useful instructions by moving from understanding, to comparing, and finally to generating the instructions.

4. Model Training

This section presents the complete training pipeline of Z-Image and Z-Image-Edit. We begin by introducing our Scalable Single-Stream Diffusion Transformer (S3-DiT) architecture (Section 4.1) and training efficiency optimizations (Section 4.2), followed by a multi-stage training process: pre-training (Section 4.3), supervised fine-tuning (Section 4.4), few-step distillation (Section 4.5), and reinforcement learning with human feedback (Section 4.6). Finally, we describe the continued training strategy for image editing capabilities (Section 4.7) and our reasoning-enhanced prompt enhancer (Section 4.8). The overall training pipeline is summarized in Figure 11. And in Figure 12, we present intermediate generation results throughout Z-Image's training process to demonstrate the benefits contributed by each stage.

4.1. Architecture

Efficiency and stability are the core objectives guiding the design of Z-Image. To achieve this, we employ the lightweight Qwen3-4B [85] as the text encoder, leveraging its bilingual proficiency to align complex instructions with visual content. For image tokenization, we utilize the Flux VAE [35] selected for its proven reconstruction quality. Exclusively for editing tasks, we augment the architecture with SigLIP 2 [69] to capture abstract visual semantics from reference images. Inspired by the scaling success of decoder-only models, we adopt a Single-Stream Multi-Modal Diffusion Transformer (MM-DiT) paradigm [18]. In this setup, text, visual semantic tokens, and VAE image tokens are concatenated at the sequence level to serve as a unified input stream, maximizing parameter efficiency compared to dual-stream approaches [18, 76]. We employ 3D Unified RoPE [58, 78] to model this mixed sequence, wherein image tokens expand across spatial dimensions and text tokens increment along the temporal dimension. Crucially, for editing tasks, the reference image tokens and target image tokens are assigned aligned spatial RoPE coordinates but are separated by a unit interval offset in the temporal dimension. Additionally, different time-conditioning values are applied to the reference and target images to distinguish between clean and noisy images.
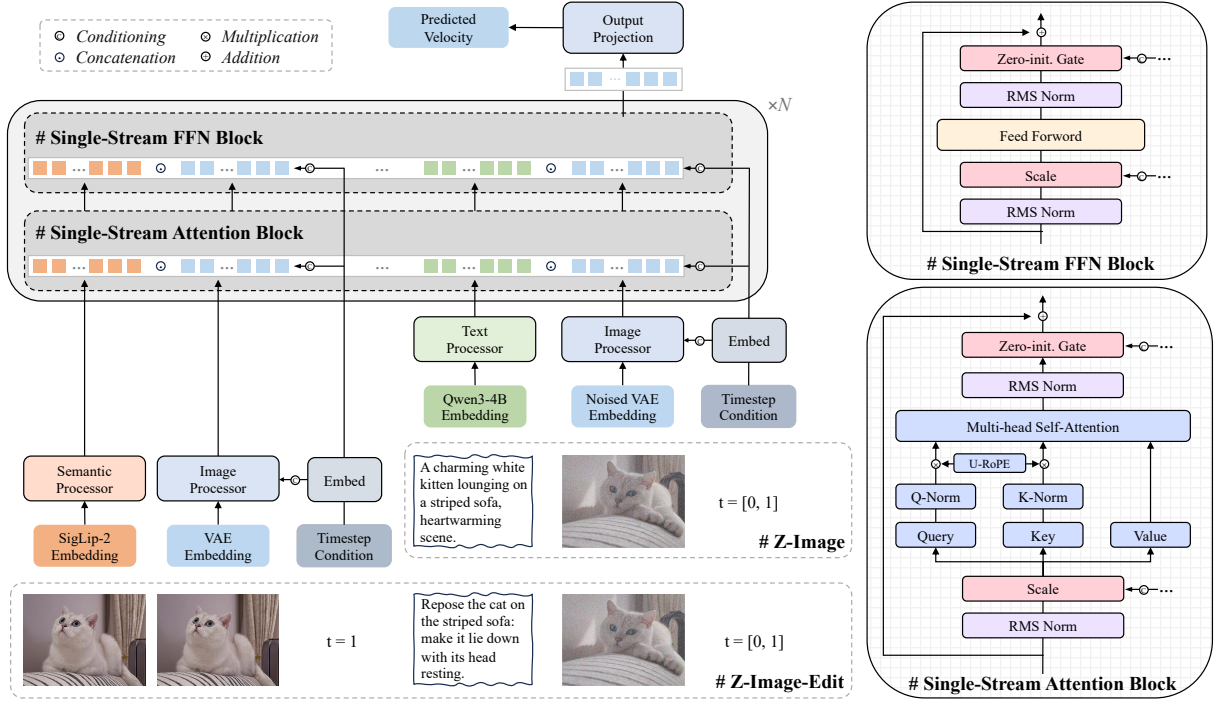
Figure 10 | Architecture overview of the Z-Image series. The S3-DiT consists of single-stream FFN blocks and single-stream attention blocks. It processes inputs from different modalities through lightweight modality-specific processors, then concatenates them into a unified input sequence. This modality-agnostic architecture maximizes cross-modal parameter reuse to ensure parameter efficiency, while providing flexible compatibility for varying input configurations in both Z-Image and Z-Image-Edit.

As illustrated in Figure 10, the specific architecture of our S3-DiT (Scalable Single-Stream DiT) commences with lightweight modality-specific processors, each composed of two transformer blocks for initial modal alignment. Subsequently, tokens enter the unified single-stream backbone. To ensure training stability, we implement QK-Norm to regulate attention activations [33, 50, 24, 53] and Sandwich-Norm to constrain signal amplitudes at the input and output of each attention / FFN blocks [16, 99]. For conditional information injection, input condition vectors are projected into scale and gate parameters to modulate the normalized inputs and outputs of both Attention and FFN layers. To reduce parameter overhead, this projection is decomposed into a low-rank pair: a shared, layer-agnostic down-projection layer followed by layer-specific up-projection layers. Finally, RMSNorm [91] is uniformly utilized for all the aforementioned normalization operations.

Table 2 | Architecture Configurations of S3-DiT.

Configuration	S3-DiT
Total Parameters	6.15B
Number of Layers	30
Hidden Dimension	3840
Number of Attention Heads	32
FFN Intermediate Dimension	10240
(d_t, d_h, d_w)	$(32, 48, 48)$

4.2. Training Efficiency Optimization

To optimize training efficiency, we implemented a multi-faceted strategy targeting both computational and memory overheads.
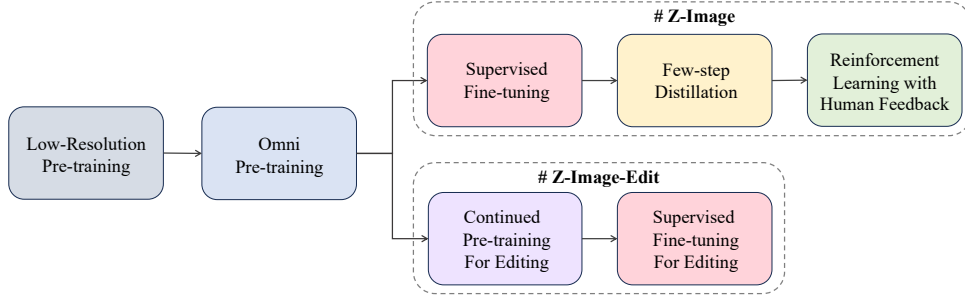
16

Figure 11 | The training pipeline of Z-Image and Z-Image-Edit. The low-resolution pre-training and omni-pre-training stages provide a suitable initialization for image generation and editing tasks, after which separate post-training processes yield the Z-Image and Z-Image-Edit models respectively.

For distributed training, we employed a hybrid parallelization strategy. We applied standard Data Parallelism (DP) to the VAE and Text Encoder, as they remain frozen and incur minimal memory footprint. In contrast, for the large DiT model, where optimizer states and gradients consume substantial memory, we utilized FSDP2 [96] to effectively shard these overheads across GPUs. Furthermore, we implemented gradient checkpointing across all DiT layers. This technique trades an acceptable increase in computational cost for significant memory savings, enabling larger batch sizes and improved overall throughput. To further accelerate computation and optimize memory usage, the DiT blocks were compiled using `torch.compile`, a just-in-time (JIT) compiler [1].

In addition to system-level optimizations, we addressed inefficiencies arising from mixed-resolution training. Grouping samples with significantly different sequence lengths into a single batch typically results in excessive padding, which significantly impedes overall training speed. To mitigate this, we designed a sequence length-aware batch construction strategy. Prior to training, we estimate the sequence length of each sample based on the resolution (height and width) recorded in the metadata. The sampler then groups samples with similar sequence lengths into the same batch, thereby minimizing computational waste. Crucially, we additionally employ a dynamic batch sizing mechanism: smaller batch sizes are assigned to long-sequence batches to prevent Out-Of-Memory (OOM) errors, while larger batch sizes are used for short sequences to avoid resource vacancy. This approach ensures maximal hardware utilization across varying resolutions.

4.3. Pre-training

Z-Image is trained using the flow matching objective [45, 48], where noised inputs are first constructed through linear interpolation between Gaussian noise x_0 and the original image x_1, i.e., $x_t = t \cdot x_1 + (1-t) \cdot x_0$. The model is then trained to predict the velocity of the vector field that defines the path between them, i.e., $v_t = x_1 - x_0$. The training objective can be formulated as:

$$\mathcal{L} = \mathbb{E}_{t,x_0,x_1,y}[\|u(x_t,y,t;\theta) - (x_1 - x_0)\|^2], \tag{1}$$

Where θ as the learnable parameters and y as the conditional embedding. Following SD3 [18], we employ the logit-normal noise sampler to concentrate the training process on intermediate timesteps. Additionally, to account for the variations in Signal-to-Noise Ratio (SNR) arising from our multi-resolution training setup, we adopt the dynamic time shifting strategy as used in Flux [35]. This ensures that the noise level is appropriately scaled for different image resolutions, leading to more effective training.

The pre-training of Z-Image can be broadly divided into two phases: low-resolution pre-training and omni-pre-training.

Low-resolution Pre-training. This phase consists of a single stage, conducted exclusively at a 256^2 resolution on the text-to-image generation task. The primary emphasis of this stage is on efficient cross-modal alignment and knowledge injection – equipping the model with the capability to generate a diverse range of concepts, styles, and compositions, which is consistent with the initial stage of conventional multi-stage training protocols. As shown in Figure 1, this phase accounts for over half of our total pre-

training compute. This allocation is based on the rationale that the majority of the model's foundational visual knowledge (*e.g.*, Chinese text rendering) is acquired during this low-resolution training stage.

Omni-pre-training. The "omni" here signifies three key aspects:

- **Arbitrary-Resolution Training**: We design an arbitrary-resolution training strategy in which the original image resolution is mapped to a predefined training resolution range through a resolution-mapping function. The model is then trained on images with diverse resolutions and aspect ratios. This enables the learning of cross-scale visual information, mitigates information loss caused by downsampling to a fixed resolution, and improves overall data efficiency.
- **Joint Text-to-Image and Image-to-Image Training**: We integrate the image-to-image task into the pre-training framework. By leveraging the substantial compute budget available during pre-training, we can effectively exploit large-scale, naturally occurring, and weakly aligned image pairs, as discussed in Section 2.5. Learning the relationships between natural image pairs provides a strong initialization for downstream tasks such as image editing. Importantly, we observe that this joint pre-training scheme does not introduce any noticeable performance degradation on the text-to-image task.
- **Multi-level and Bilingual Caption Training**: It is widely recognized that high-quality captions are crucial for training text-to-image models [4]. To ensure both bilingual understanding and strong native prompt-following capability, we employ Z-Captioner to generate bilingual, multi-level synthetic captions (including long, medium, and short descriptions, as well as tags and simulated user prompts). In addition, the original textual metadata associated with each image is incorporated with a small probability to further enhance the model's acquisition of world knowledge. The use of captions at different granularities and from diverse perspectives provides broad mode coverage, which is beneficial for subsequent stages of training. Moreover, for image-to-image tasks, we randomly sample either the target image's caption or the pairwise difference caption with a certain probability, corresponding to reference-guided image generation and multi-task image editing, respectively.

Working with our data infrastructure, the omni-pre-training phase is conducted in multiple stages. Upon completion of the final stage, the model becomes capable of generating images at arbitrary resolutions up to the 1k-1.5k range and can condition its output on both image and text inputs. This provides a suitable starting point for the subsequent training of Z-Image and Z-Image-Edit.

4.4. Supervised Fine-Tuning (SFT)

Distribution Narrowing via High-Quality Alignment. While the omni-pre-training stage establishes broad world understanding and mode coverage, the resulting distribution inevitably exhibits high variance, reflecting the noisy nature of web-scale data. Consequently, the primary objective of Supervised Fine-Tuning (SFT) is not merely to correct local artifacts, but to narrow the generation distribution **towards a focused, high-fidelity sub-manifold** [67]. This phase aims for rapid convergence to a fixed distribution characterized by consistent visual aesthetics and precise instruction following. To achieve this, we transition from the noisy supervision of pre-training to a curriculum dominated by highly curated images filtering by our data infrastructure and *super detailed, grounded captions*. This rigorous supervision acts as an anchor, forcing the model to discard low-quality modes (*e.g.*, unstable stylization or inconsistent rendering) and align strictly with detailed textual descriptions, shifting the model from a diversity-maximizing regime to a quality-maximizing operating point.

Concept Balancing with Tagged Resampling. A critical challenge in narrowing the distribution is the risk of catastrophic forgetting, particularly for long-tail concepts that are prone to being overshadowed by dominant modes during convergence. To address this, we enforce strict **class balancing** throughout the SFT phase. We employ a dynamic resampling strategy guided by world knowledge topological graph in Section 2. Specifically, we maintain a target prior over concepts and utilize BM25-based retrieval to compute rarity scores for training samples on the fly. Mini-batches are constructed by up-weighting under-represented concepts – such as rare entities or specific artistic styles – while down-weighting over-represented ones. This mechanism ensures that while the model converges to the target high-quality

Prompt	(a) Pre-train	(b) SFT	(c) PE	(d) FSD	(e) RLHF

At a dining table, an elderly grandmother and her young granddaughter share a quiet moment, looking through an old photo album. The grandmother, whose expression is one of wise and patient kindness, smiles gently. The warm lamplight catches the scene, casting a soft glow on both their faces, which are mirrored in the polished, dark wood surface of the table.

画面中，一只拟人化的红熊猫身着蒸汽朋克飞行装束，头戴护目镜，正举起一只黄铜和紫铜制成的假肢，自信地敬礼。他透过齐柏林飞艇光滑的弧形黄铜舷窗，看着自己的倒影，脸上带着骄傲的笑容，倒影也模仿着他的动作。

A young man gently holds a woman's hand, helping her maintain balance as she gracefully twirls on the wet, reflective cobblestones of a Parisian street at night. They are both laughing, caught in a moment of spontaneous joy. The polished stone street, slick with recent rain, acts as a dark, distorted mirror, reflecting their dancing figures and the warm glow of the streetlights around them.

Figure 12 | Intermediate generation results throughout Z-Image-Turbo's training process, echoing our analysis of each stage's contribution.

distribution, the marginal distribution over concepts remains uniform, effectively preserving the semantic diversity of the pre-trained model.

Robustness via Model Merging. Despite balanced training, SFT on specific high-quality datasets can introduce subtle biases or trade-offs between capabilities (*e.g.*, photorealism vs. stylistic flexibility). To achieve a Pareto-optimal solution without complex inference routing, we employ **Model Merging** [75, 93] as the final refinement step. We fine-tune multiple SFT variants initialized from the same backbone, each slightly biased towards different capability dimensions (*e.g.*, strict instruction following or aesthetic rendering). We then perform a linear interpolation of their weights in the parameter space: $\theta_{\text{final}} = \sum_i \alpha_i \theta_i$. This lightweight merging strategy effectively smooths the loss landscape, neutralizing individual biases and resulting in a final model that exhibits superior stability and robustness across diverse prompts compared to any single SFT checkpoint.

4.5. Few-Step Distillation

The goal of the Few-Step Distillation stage is to reduce the inference time of our foundational SFT model, achieving the efficiency demanded by real-world applications and large-scale deployment. While our 6B foundational model represents a significant leap in efficiency compared to larger counterparts, the inference cost remains non-negligible. Due to the inherent iterative nature of diffusion models, our standard SFT model requires approximately 100 Number of Function Evaluations (NFEs) to generate high-quality samples using Classifier-Free Guidance (CFG) [29]. To bridge the gap between generation quality and interactive latency, we implemented a few-step distillation strategy.

Fundamentally, the distillation process involves teaching a student model to mimic the teacher's denoising dynamics across fewer timesteps along its sampling trajectory. The core challenge lies in reducing the inherent uncertainty of this trajectory, allowing the student to "collapse" its probabilistic path into a deterministic and highly efficient inference process. Therefore, the key to enable a stable few-step integrator is to meticulously control the distillation process. We initially selected the Distribution Matching Distillation (DMD) [88, 89] paradigm due to its promising performance in academic works. However, in practice, we encountered persistent artifacts such as the loss of high-frequency details and noticeable color shifts – issues that have been increasingly documented by the community. These observations signaled a need for algorithmic refinement. Through a deeper exploration of the distillation mechanism, we gained new

insights into the underlying dynamics of DMD, leading to two key technical advancements: *Decoupled DMD* [46] and *DMDR* [32]. We refer interested readers to the respective academic papers for full technical details. Below, we introduce the practical application of these techniques in building **Z-Image-Turbo**.

Prompt	(a) SFT	(b) DMD	(c) D-DMD	(d) D-DMD+DMDR
黄昏时分，在一间堆满奇珍异宝的杂乱古董店里，一位神情若有所思的年轻女子轻轻抚摸着一个古老硕大天球饱经风霜的表面。 At dusk, in a cluttered antique shop filled with curiosities, a young woman with a thoughtful expression gently touches the weathered surface of a large, ancient celestial globe.				
一只张开的手的特写，手心向上，掌纹和指关节的细节清晰可见。 A close-up of an open hand, palm up, with the details of the palm lines and knuckles clearly visible.				
一幅达芬奇风格的文艺复兴时期油画，描绘了圣母玛利亚怀抱圣子耶稣，人物构成了稳定的金字塔形，背景是朦胧的风景。 A Renaissance oil painting in the style of Leonardo da Vinci depicts the Virgin Mary holding the infant Jesus. The figures form a stable pyramid shape against a hazy landscape background.				

Figure 13 | Few-Step Distillation visualization results across different distillation strategies: (a) the original SFT model; (b) Standard DMD; (c) *Decoupled DMD (D-DMD)*; and (d) *D-DMD+DMDR* (**Z-Image-Turbo**). The proposed approach achieves real-time 8-step inference while attaining superior perceived quality and aesthetic appeal.

4.5.1. Decoupled DMD: Resolving Detail and Color Degradation

Our investigation revealed a core insight: the effectiveness of existing DMD methods is not a monolithic phenomenon but the result of two independent, collaborating mechanisms:

- **CFG-Augmentation (CA):** This acts as the primary engine driving the distillation process, efficiently building up the few-step generation capabilities of the student model. Despite its dominant role, this factor has been largely overlooked in previous literature.
- **Distribution Matching (DM):** This functions primarily as a powerful regularizer, ensuring the stability of the training process and removing the emerging artifacts.

By recognizing and decoupling these two mechanisms, we were able to study and optimize them in isolation. This motivation led to the development of *Decoupled DMD*, an improved distillation framework that features a decoupled application of renoising schedules tailored specifically for the CA and DM terms.

In practice, *Decoupled DMD* effectively addresses the pain points of traditional DMD, ensuring sharp detail preservation and color fidelity. Notably, the resulting distilled model not only matches the original multi-step teacher but even surpasses it in terms of photorealism and visual impact.

4.5.2. DMDR: Enhancing Capacity with RL and Regularization

To further push the performance boundaries of our few-step model, we incorporate Reinforcement Learning (RL) into the few-step distillation process. Applying RL to generative models typically faces the risk of "reward hacking", where the model exploits the reward function to generate high-scoring but visually nonsensical images. To mitigate this, external regularization is usually required.

Our insight from Decoupled DMD provides a natural solution: since we established that the Distribution Matching (DM) term functions as a high-quality regularizer, it can be organically combined with RL objectives. This synthesis gives rise to *DMDR* (Distribution Matching Distillation meets Reinforcement Learning) [32]. In this framework, RL unlocks the student model's capacity to align with human preferences, while the DM term acts as a robust constraint, effectively preventing reward hacking. This synergy allows **Z-Image-Turbo** to achieve superior aesthetic alignment and semantic faithfulness while maintaining strict generative stability.

4.5.3. Results and Analysis

The efficacy of our *Decoupled DMD* and *DMDR* distillation strategy is visualized in Figure 13. The original SFT model (a) sets a high baseline but suffers from high latency. Standard DMD (b), while fast, exhibits characteristic degradation: blurred textures and shifted color tones. Our *Decoupled DMD* (c) successfully resolves these artifacts, restoring sharp details and accurate colors. Finally, **Z-Image-Turbo** (d), refined via a combination of *Decoupled DMD* and *DMDR*, represents the optimal convergence of speed and quality. It achieves 8-step inference that is not only indistinguishable from the 100-step teacher but frequently surpasses it in perceived quality and aesthetic appeal. In summary, our Few-Step Distillation framework resolves the long-standing tension between inference speed and visual fidelity.

4.6. Reinforcement Learning with Human Feedback (RLHF)

Following the previous stages, the model has acquired strong foundational capabilities but may still exhibit inconsistencies in aligning with nuanced human preferences. To bridge this gap, we introduce a comprehensive post-training framework leveraging Reinforcement Learning with Human Feedback (RLHF). This framework hinges on a powerful, multi-dimensional reward model, which provides targeted feedback for online optimization. Guided by these feedback signals, our approach is structured into two sequential stages: an initial offline alignment phase using Direct Preference Optimization (DPO) [59], followed by an online refinement phase with Group Relative Policy Optimization (GRPO) [66]. This two-stage strategy allows us to first efficiently instill robust adherence to objective standards and then leverage the fine-grained signals from our reward model for optimizing more subjective qualities. As illustrated in Figure 14, this comprehensive process yields substantial improvements in photorealism, aesthetic quality, and instruction following.

4.6.1. Reward Annotation and Training

As an indispensable and critical component of the RLHF pipeline, our reward model is designed to evaluate the model's performance along three key dimensions: instruction-following capability, AI-Content Detection perception, and aesthetic quality. The reward model is then trained specifically to provide targeted feedback along these axes. For instruction following, we perform syntactic and semantic decomposition of the prompt into a structured hierarchy that includes (i) core subject entities, (ii) attribute specifications, (iii) action or interaction requirements, (iv) spatial or compositional constraints, and (v) stylistic or rendering conditions. During annotation, human raters simply click on the elements that are not satisfied by the model's output. We then compute the ratio of satisfied elements to obtain the final instruction-following score, which is used as the target reward.

4.6.2. Stage 1: Offline Alignment with DPO on Objective Dimensions

While manually curating preference pairs for DPO is feasible for capturing human aesthetic judgments, scaling this process to a large, high-quality dataset presents a significant bottleneck in real practice. Sourcing consistently informative preference pairs across subjective dimensions (*e.g.*, aesthetics, style)

近景特写一只人类异色眼眸，蓝色瞳孔

在一个阳光明媚的夏日，一个美丽而天真的18岁女孩在小溪里玩水，笑容灿烂

香蕉长出猴子的脑袋

一个狮子骑的袋鼠在喜马拉雅山脉与运动员奔跑的画面

Figure 14 | **Visual comparison between Few-Step Distillation (FSD, top row) and RLHF (bottom row).** Building upon the strong foundation of the FSD model, RLHF further enhances **photorealism**, **aesthetic quality**, and **instruction following**.

is slow and requires extensive expert annotation. To address this scalability challenge and enhance annotation efficiency, our DPO strategy pivots to focus exclusively on objective, verifiable dimensions.

These dimensions, such as **text rendering** and **object counting**, offer clear and binary correctness criteria that are highly amenable to automated evaluation by modern Vision-Language Models (VLMs). For instance, given a prompt requiring specific text, an image with accurately rendered characters is designated as the positive sample ('chosen'), while an image with typographical errors becomes the negative sample ('rejected'). We leverage VLMs to programmatically generate a large corpus of such candidate preference pairs. This VLM-generated dataset is then subjected to a streamlined human verification and cleaning process, ensuring high fidelity. This hybrid VLM-human pipeline dramatically increases annotation throughput and consistency compared to purely human manual curation.

Furthermore, to smooth the learning curve, we implement a curriculum learning strategy for DPO training. The process begins with prompts of low complexity (*e.g.*, rendering a single word, generating a small number of objects) and progressively advances to more challenging instructions involving multiple elements, complex layouts, or difficult styles. During this process, we also optimized our pair selection strategy. We observed that DPO's convergence is sensitive to the differentiation between positive and negative samples. To maximize training efficiency, our curriculum initially prioritizes pairs with moderate differentiation and gradually introduces more challenging pairs exhibiting larger or more subtle differences, which we found accelerate convergence and improve the final performance.

4.6.3. Stage 2: Online Refinement with GRPO

Building upon the robust foundation established by DPO, the second stage employs online reinforcement learning with GRPO. Guided by our reward model, this stage is designed to significantly enhance the model's capability for **photorealistic image generation**, alongside improving **aesthetic quality** and nuanced **instruction-following**.

During the GRPO training loop, we compute a composite advantage function by aggregating the scores from our reward model (*e.g.*, realism, aesthetics, instruction following, etc.). This multi-faceted feedback mechanism enables targeted, fine-grained optimization [84]. By providing distinct signals for different aspects of the generation, GRPO can simultaneously enhance photorealistic image generation, aesthetic quality, improve semantic accuracy, and reduce undesirable artifacts. This integrated approach proved to be significantly more effective than optimizing against a single reward, allowing the model to achieve a

better balance across multiple, often competing, quality dimensions.

4.7. Continued Training for Image Editing

Starting from the base model, the continued pre-training for image editing consists of two stages, as shown in Figure 10. In the continued pre-training stage, we train the model with our constructed editing pairs (see Section 2.5), together with our text-to-image SFT data to ensure high image quality. We first train the whole amount of editing data in resolution of 512^2 for a few thousand steps for quick adaptation to editing tasks, and then increase the image resolution to 1024^2 for high generation quality. Because image editing data pairs are expensive and difficult to acquire, their total volume is significantly smaller and far less diverse than that of text-to-image data. Therefore, we suggest a relatively higher ratio of text-to-image data (*e.g.*, text-to-image:image-to-image = 4 : 1) to avoid performance degradation during training.

In the following SFT stage, a task-balanced, high-quality subset of the training data is manually constructed to further improve the model's overall performance, especially its instruction following ability. However, synthetic data (*e.g.*, the rendered text data for text editing), though easy-to-acquire and guaranteed to be 100% accurate in terms of instruction following, are far from the distribution of real-world user input, and are thus heavily downsampled in this final training stage.

4.8. Prompt Enhancer with Reasoning Chain

Figure 15 | PE visualization. We compare generation results between PE without reasoning (middle column) and PE with reasoning (right column). As shown in the top row, the reasoning chain enables the model to decipher raw coordinates into a specific scenic context (*e.g.*, West Lake) rather than simply rendering the coordinate text. In the second row, the reasoning module plans detailed steps for "brewing Pu-erh tea," allowing the model to generate specific illustrations for each step instead of a generic list. This demonstrates that the reasoning chain effectively injects world knowledge and provides fine-grained content planning for complex user prompts.

Due to limited model size (6B parameters), Z-Image exhibits limitations in world knowledge, intent

understanding, and complex reasoning. However, it serves as a powerful text decoder capable of translating detailed prompts into realistic images. To address the cognitive gaps, we propose equipping Z-Image with a Prompt Enhancer (PE), powered by system prompt and a pretrained VLM, to improve its reasoning and knowledge capabilities.

Distinct from other methods, we keep the large VLM fixed during alignment. Instead, we process all input prompts (and input images for Z-Image-Edit) through our PE model during the Supervised Fine-Tuning (SFT) stage. This strategy ensures that Z-Image aligns effectively with the Prompt Enhancer during SFT. Furthermore, we identify the structured reasoning chain as a key factor for injecting reasoning and world knowledge. As shown in Figure 15, without reasoning, the PE merely renders coordinate text onto the image when given geolocation data; with reasoning, it infers the location (*e.g.*, West Lake) to generate the correct scene. Similarly, in generating journal-style instructions, the lack of reasoning leads to monotonous outputs, whereas the reasoning-enhanced model enriches the result by generating specific illustrations for each step.

5. Performance Evaluation

5.1. Elo-based Human Preference Evaluation

Table 3 | Elo rankings of Text-to-Image models from AI Arena. Z-Image-Turbo achieves 4th globally and 1st among open-source models.

Rank	Model Name	Company	Type	95% CI	Elo Score	Win Rate
1	Imagen 4 Ultra Preview 0606	Google	Closed-source	-16/+16	1048	48%
2	gemini-2.5-flash-image-preview	Google	Closed-source	-16/+14	1046	47%
3	Seedream 4.0	ByteDance	Closed-source	-17/+16	1039	46%
4	**Z-Image-Turbo**	Alibaba	Open-source (6B)	-15/+17	1025	45%
5	Seedream 3.0	ByteDance	Closed-source	-15/+19	1012	41%
6	Qwen-Image	Alibaba	Open-source (20B)	-16/+16	1008	41%
7	GPT Image 1	OpenAI	Closed-source	-14/+17	986	38%
8	FLUX.1 Kontext Pro	Black Forest Labs	Closed-source	-15/+14	950	32%
9	Ideogram 3.0	Ideogram	Closed-source	-15/+16	936	29%

To rigorously benchmark Z-Image-Turbo's capabilities against the competitive landscape of generative models, we participated in AI Arena[2], a public-facing, independent benchmarking platform powered by large-scale human judgment. Unlike automated metrics that frequently misalign with human perception, AI Arena provides dynamic and unbiased model rankings based on thousands of pairwise comparisons, making it an ideal venue for objective performance assessment.

The evaluation protocol is built upon the Elo rating system – a well-established method for ranking competitors based on head-to-head outcomes. In each round, two images generated from the same text prompt by different models are displayed side-by-side with identities hidden. Evaluators are asked to select the image they perceive as superior in terms of visual coherence, detail rendering, prompt alignment, and artistic quality. Each vote updates the global Elo leaderboard dynamically, ensuring that rankings reflect collective human judgment over time.

In the evaluation[3], Z-Image-Turbo – our high-efficiency diffusion architecture with 6B parameters and low inference cost of 8 NFEs – competed against 8 leading models, including top-tier closed-source systems: Imagen 4 Ultra Preview 0606 [26], gemini-2.5-flash-image-preview [25] (both Google), Seedream 4.0 (ByteDance) [64], GPT Image 1 [High] (OpenAI) [55], FLUX.1 Kontext [Pro] (Black Forest Labs) [37], and Ideogram 3.0 [31], along with the open-source baseline Qwen-Image [76].

As shown in Table 3, Z-Image-Turbo attained an Elo score of 1025, achieving 4th place globally and ranking 1st among open models on the leaderboard. Despite its efficient design, it outperforms multiple industry-leading closed-source models and surpasses Qwen-Image while operating at significantly lower computational cost. With a 45% win rate across all matchups, Z-Image-Turbo demonstrates

---

[2]https://aiarena.alibaba-inc.com/corpora/arena/leaderboard?arenaType=T2I
[3]Snapshot taken on November 26, 2025

strong competitiveness against the entire field, including top-tier proprietary systems. This high level of performance positions it firmly within the elite tier of generative models, affirming its status as one of the most advanced open-source systems available.

These results establish Z-Image-Turbo as one of the leading open text-to-image models in terms of both quality and efficiency. More than a high-performing generator, it represents a new baseline for efficiency-oriented architecture design, demonstrating that compact models can achieve elite-level performance without compromising usability. This combination of speed, fidelity, and openness enables deployment in resource-constrained environments, interactive applications, and community-driven innovation.

5.2. Quantitative Evaluation

To comprehensively evaluate the generation and editing capabilities of **Z-Image** and its variants (**Z-Image-Turbo** and **Z-Image-Edit**), we conducted extensive experiments across multiple authoritative benchmarks. These evaluations cover general image generation, fine-grained instruction following, text rendering in both English and Chinese, and instruction-based image editing.

5.2.1. Text-to-Image Generation

CVTG-2K. To evaluate our model's performance on text rendering tasks, we conduct quantitative experiments on the CVTG-2K benchmark [17]. CVTG-2K is a specialized benchmark designed for Complex Visual Text Generation, encompassing diverse scenarios with varying numbers of text regions. As presented in Table 4, our model achieves superior performance on CVTG-2K across all evaluation metrics. Specifically, Z-Image attains the highest average Word Accuracy score of 0.8671, outperforming competitive baselines such as GPT-Image-1 [55] (0.8569) and Qwen-Image [76] (0.8288). Notably, our model demonstrates robust performance across varying levels of complexity, maintaining consistent accuracy even as the number of text regions increases from 2 to 5. Furthermore, Z-Image-Turbo, our efficient variant, achieves the highest CLIP Score of 0.8048 among all models while maintaining competitive text accuracy (0.8585 average Word Accuracy), striking an effective balance between generation quality and inference efficiency. These results demonstrate the effectiveness of our approach in complex visual text generation scenarios.

Table 4 | Quantitative evaluation results of English text rendering on CVTG-2K [17].

Rank	Model	NED	CLIPScore	Word Accuracy				
				2 regions	3 regions	4 regions	5 regions	average↑
1	**Z-Image**	0.9367	0.7969	**0.9006**	**0.8722**	0.8652	**0.8512**	**0.8671**
2	**Z-Image-Turbo**	0.9281	**0.8048**	0.8872	0.8662	0.8628	0.8347	0.8585
3	GPT Image 1 [High] [55]	**0.9478**	0.7982	0.8779	0.8659	**0.8731**	0.8218	0.8569
4	Qwen-Image [76]	0.9116	0.8017	0.8370	0.8364	0.8313	0.8158	0.8288
5	TextCrafter [17]	0.8679	0.7868	0.7628	0.7628	0.7406	0.6977	0.7370
6	SD3.5 Large [18]	0.8470	0.7797	0.7293	0.6825	0.6574	0.5940	0.6548
7	Seedream 3.0 [21]	0.8537	0.7821	0.6282	0.5962	0.6043	0.5610	0.5924
8	FLUX.1 [dev] [37]	0.6879	0.7401	0.6089	0.5531	0.4661	0.4316	0.4965
9	3DIS [98]	0.6505	0.7767	0.4495	0.3959	0.3880	0.3303	0.3813
10	RAG-Diffusion [40]	0.4498	0.7797	0.4388	0.3316	0.2116	0.1910	0.2648
11	TextDiffuser-2 [10]	0.4353	0.6765	0.5322	0.3255	0.1787	0.0809	0.2326
12	AnyText [70]	0.4675	0.7432	0.0513	0.1739	0.1948	0.2249	0.1804

LongText-Bench. To further assess our model's capability in rendering longer texts, we evaluate its performance on LongText-Bench [22], a specialized benchmark focusing on evaluating the performance of rendering longer texts in both English and Chinese. As shown in Table 5, our models demonstrate strong and consistent performance across both language settings. On LongText-Bench-EN, Z-Image achieves a competitive score of 0.935, ranking third among all evaluated models, while on LongText-Bench-ZH, it attains a score of 0.936, securing second place. Z-Image-Turbo also delivers impressive results, scoring 0.917 on the English benchmark and 0.926 on the Chinese benchmark, demonstrating strong efficiency-performance trade-offs. This consistent performance across both languages highlights our model's robust bilingual text rendering capability.

Table 5 | Quantitative evaluation results on LongText-Bench [22].

Rank	Model	LongText-Bench-EN↑	LongText-Bench-ZH↑
1	Qwen-Image [76]	0.943	**0.946**
2	**Z-Image**	0.935	0.936
3	**Z-Image-Turbo**	0.917	0.926
4	Seedream 3.0 [21]	0.896	0.878
5	X-Omni [22]	0.900	0.814
6	GPT Image 1 [High] [55]	**0.956**	0.619
7	Kolors 2.0 [34]	0.258	0.329
8	BAGEL [15]	0.373	0.310
9	OmniGen2 [78]	0.561	0.059
10	HiDream-I1-Full [7]	0.543	0.024
11	BLIP3-o [11]	0.021	0.018
12	Janus-Pro [14]	0.019	0.006
13	FLUX.1 [Dev] [37]	0.607	0.005

OneIG. We utilize the OneIG benchmark [9] to assess fine-grained alignment. As reported in Tables 6 and 7, Z-Image achieves the highest overall score (0.546) on the English track, surpassing Qwen-Image (0.539) and GPT Image 1 [High] (0.533). Notably, Z-Image sets a new state-of-the-art in text rendering reliability with an English *Text* score of 0.987 and a Chinese *Text* score of 0.988, significantly outperforming competitors. On the Chinese track, Z-Image ranks second overall (0.535), confirming its multi-lingual robustness. Additionally, our distilled version, Z-Image-Turbo, demonstrates impressive efficiency, maintaining strong performance with only a marginal decrease compared to the base model.

Table 6 | Quantitative evaluation results on OneIG-EN [9]. The overall score is the average of the five dimensions.

Rank	Model	Alignment	Text	Reasoning	Style	Diversity	Overall↑
1	**Z-Image**	0.881	0.987	0.280	0.387	0.194	**0.546**
2	Qwen-Image [76]	**0.882**	0.891	0.306	0.418	0.197	0.539
3	GPT Image 1 [High] [55]	0.851	0.857	**0.345**	**0.462**	0.151	0.533
4	Seedream 3.0 [21]	0.818	0.865	0.275	0.413	0.277	0.530
5	**Z-Image-Turbo**	0.840	**0.994**	0.298	0.368	0.139	0.528
6	Imagen 4 [26]	0.857	0.805	0.338	0.377	0.199	0.515
7	Recraft V3 [61]	0.810	0.795	0.323	0.378	0.205	0.502
8	HiDream-I1-Full [7]	0.829	0.707	0.317	0.347	0.186	0.477
9	OmniGen2 [78]	0.804	0.680	0.271	0.377	0.242	0.475
10	SD3.5 Large [18]	0.809	0.629	0.294	0.353	0.225	0.462
11	CogView4 [97]	0.786	0.641	0.246	0.353	0.205	0.446
12	FLUX.1 [Dev] [37]	0.786	0.523	0.253	0.368	0.238	0.434
13	Kolors 2.0 [34]	0.820	0.427	0.262	0.360	0.300	0.434
14	Imagen 3 [3]	0.843	0.343	0.313	0.359	0.188	0.409
15	BAGEL [15]	0.769	0.244	0.173	0.367	0.251	0.361
16	Lumina-Image 2.0 [58]	0.819	0.106	0.270	0.354	0.216	0.353
17	SANA-1.5-4.8B [81]	0.765	0.069	0.217	0.401	0.216	0.334
18	SANA-1.5-1.6B [81]	0.762	0.054	0.209	0.387	0.222	0.327
19	BAGEL+CoT [15]	0.793	0.020	0.206	0.390	0.209	0.324
20	SD 1.5 [63]	0.565	0.010	0.207	0.383	**0.429**	0.319
21	SDXL [57]	0.688	0.029	0.237	0.332	0.296	0.316
22	Show-o2-7B [83]	0.817	0.002	0.226	0.317	0.177	0.308
23	BLIP3-o [11]	0.711	0.013	0.223	0.361	0.229	0.307
24	Show-o2-1.5B [83]	0.798	0.002	0.219	0.317	0.186	0.304
25	Janus-Pro [14]	0.553	0.001	0.139	0.276	0.365	0.267

GenEval. As shown in Table 8, we evaluate object-centric generation using GenEval [23]. Z-Image achieves an overall score of 0.84, securing a three-way tie for second place alongside Seedream 3.0 [21] and GPT Image 1 [High] [55], trailing only Qwen-Image [76] (0.87). Notably, Z-Image-Turbo delivers highly competitive performance with an overall score of 0.82, maintaining only a 2-point gap from the base model. These results indicate that our model possesses a robust capability for generating accurate and distinct entities.

Table 7 | Quantitative evaluation results on OneIG-ZH [9]. The overall score is the average of the five dimensions.

Rank	Model	Alignment	Text	Reasoning	Style	Diversity	Overall↑
1	Qwen-Image [76]	**0.825**	0.963	0.267	0.405	0.279	**0.548**
2	**Z-Image**	0.793	**0.988**	0.266	0.386	0.243	0.535
3	Seedream 3.0 [21]	0.793	0.928	0.281	0.397	0.243	0.528
4	**Z-Image-Turbo**	0.782	0.982	0.276	0.361	0.134	0.507
5	GPT Image 1 [High] [55]	0.812	0.650	**0.300**	**0.449**	0.159	0.474
6	Kolors 2.0 [34]	0.738	0.502	0.226	0.331	0.333	0.426
7	BAGEL [15]	0.672	0.365	0.186	0.357	0.268	0.370
8	Cogview4 [97]	0.700	0.193	0.236	0.348	0.214	0.338
9	HiDream-I1-Full [7]	0.620	0.205	0.256	0.304	0.300	0.337
10	Lumina-Image 2.0 [58]	0.731	0.136	0.221	0.343	0.240	0.334
11	BAGEL+CoT [15]	0.719	0.127	0.219	0.385	0.197	0.329
12	BLIP3-o [11]	0.608	0.092	0.213	0.369	0.233	0.303
13	Janus-Pro [14]	0.324	0.148	0.104	0.264	**0.358**	0.240

Table 8 | Quantitative Evaluation results on GenEval [23].

Rank	Model	Single Object	Two Object	Counting	Colors	Position	Attribute Binding	Overall↑
1	Qwen-Image [76]	0.99	0.92	0.89	0.88	0.76	0.77	**0.87**
2	**Z-Image**	1.00	0.94	0.78	**0.93**	0.62	0.77	0.84
2	Seedream 3.0 [21]	0.99	0.96	**0.91**	**0.93**	0.47	**0.80**	0.84
2	GPT Image 1 [High] [55]	0.99	0.92	0.85	0.92	0.75	0.61	0.84
5	HiDream-I1-Full [7]	1.00	**0.98**	0.79	0.91	0.60	0.72	0.83
6	**Z-Image-Turbo**	1.00	0.95	0.77	0.89	0.65	0.68	0.82
7	Janus-Pro-7B [14]	0.99	0.89	0.59	0.90	**0.79**	0.66	0.80
8	Lumina-Image 2.0 [58]	-	0.87	0.67	-	-	0.62	0.73
9	SD3.5-Large [18]	0.98	0.89	0.73	0.83	0.34	0.47	0.71
10	FLUX.1 [Dev] [37]	0.98	0.81	0.74	0.79	0.22	0.45	0.66
11	JanusFlow [51]	0.97	0.59	0.45	0.83	0.53	0.42	0.63
12	SD3 Medium [18]	0.98	0.74	0.63	0.67	0.34	0.36	0.62
13	Emu3-Gen [72]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
14	Show-o [82]	0.95	0.52	0.49	0.82	0.11	0.28	0.53
15	PixArt-α [13]	0.98	0.50	0.44	0.80	0.08	0.07	0.48

DPG-Bench. Table 9 presents the comparison on the DPG-Bench benchmark [30], which evaluates the ability of prompt following in dense prompts. Z-Image achieves a strong global performance, ranking third overall with a score of 88.14, closely trailing Seedream 3.0 [21] and Qwen-Image [76]. Notably, our model demonstrates robust performance in the *Attribute* dimension (93.16), surpassing the leading Qwen-Image (92.02) and Seedream 3.0 (91.36). Furthermore, our 8-step distillation model (Z-Image-Turbo), maintains competitive performance while achieving high efficiency.

TIIF. Table 10 details the performance on the TIIF benchmark testmini [74], which systematically evaluates instruction-following capabilities. Z-Image and Z-Image-Turbo achieve the 4th and 5th ranks, respectively. These results demonstrate that both the base and distilled versions possess exceptional capabilities in interpreting and executing complex user instructions across diverse categories.

PRISM-Bench. We evaluate our models on PRISM-Bench [19], a VLM-powered benchmark assessing reasoning and aesthetics across seven tracks. On the English track (Table 11), Z-Image-Turbo achieves the 3rd rank (77.4), outperforming the base model and Qwen-Image, which highlights its superior efficiency and generation quality. On the Chinese track (Table 12), Z-Image ranks 2nd (75.3), demonstrating robust multi-lingual performance with exceptional scores in *Text Rendering* (83.4) and *Composition* (88.6).

Table 9 | Quantitative evaluation results on DPG [30].

Rank	Model	Global	Entity	Attribute	Relation	Other	Overall↑
1	Qwen-Image [76]	91.32	91.56	92.02	94.31	**92.73**	**88.32**
2	Seedream 3.0 [21]	**94.31**	**92.65**	91.36	92.78	88.24	88.27
3	**Z-Image**	93.39	91.22	**93.16**	92.22	91.52	88.14
4	Lumina-Image 2.0 [58]	-	91.97	90.20	**94.85**	-	87.20
5	HiDream-I1-Full [7]	76.44	90.22	89.48	93.74	91.83	85.89
6	GPT Image 1 [High] [55]	88.89	88.94	89.84	92.63	90.96	85.15
7	**Z-Image-Turbo**	91.29	89.59	90.14	92.16	88.68	84.86
8	Janus-Pro-7B [14]	86.90	88.90	89.40	89.32	89.48	84.19
9	SD3 Medium [18]	87.90	91.01	88.83	80.70	88.68	84.08
10	FLUX.1 [Dev] [37]	74.35	90.00	88.96	90.87	88.33	83.84
11	DALL-E 3 [4]	90.97	89.61	88.39	90.58	89.83	83.50
12	Janus-Pro-1B [14]	87.58	88.63	88.17	88.98	88.30	82.63
13	Emu3-Gen [72]	85.21	86.68	86.84	90.22	83.15	80.60
14	PixArt-Σ [12]	86.89	82.89	88.94	86.59	87.68	80.54
15	Janus [77]	82.33	87.38	87.70	85.46	86.41	79.68
16	Hunyuan-DiT [41]	84.59	80.59	88.01	74.36	86.41	78.87
17	Playground v2.5 [39]	83.06	82.59	81.20	84.08	83.50	75.47
18	SDXL [57]	83.27	82.43	80.91	86.76	80.41	74.65
19	Lumina-Next [99]	82.82	88.65	86.44	80.53	81.82	74.63
20	PixArt-α [13]	74.97	79.32	78.60	82.57	76.96	71.11
21	SD1.5 [63]	74.63	74.23	75.39	73.49	67.81	63.18

Table 10 | Quantitative evaluation results on TIIF Bench testmini [74].

Rank	Model	Overall↑		Basic Following								Advanced Following													Designer	
				Avg		Attribute		Relation		Reasoning		Avg		Attr.+Rela.		Attr.+Reas.		Rela.+Reas.		Style		Text		Real World	
		short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long
1	GPT Image 1 [High] [55]	89.15	88.29	90.75	89.66	91.33	87.08	84.57	84.57	96.32	97.32	88.55	88.35	87.07	89.44	87.22	83.96	85.59	83.21	90.00	93.33	89.83	86.83	89.73	93.46
2	Qwen-Image [76]	86.14	86.83	90.18	87.22	90.50	91.50	88.22	90.78	79.81	79.38	79.30	80.88	79.21	78.94	78.85	81.69	75.57	78.59	100.00	100.00	92.76	89.14	90.30	91.42
3	Seedream 3.0 [21]	86.02	84.31	87.07	84.93	90.50	90.00	89.85	85.94	80.86	78.86	79.16	80.60	79.76	81.82	77.23	78.85	75.64	78.64	100.00	93.33	97.17	87.78	83.21	83.58
4	**Z-Image**	80.20	83.04	78.36	82.79	79.50	86.50	80.45	79.94	75.13	81.94	72.89	77.02	72.91	77.56	66.99	73.82	73.89	75.62	90.00	93.33	94.84	93.21	88.06	85.45
5	**Z-Image-Turbo**	77.73	80.05	81.85	81.59	86.50	87.00	82.88	79.99	76.17	77.77	68.32	74.69	72.04	75.24	60.22	73.33	68.90	71.92	83.33	93.33	83.71	84.62	85.82	77.24
6	DALL-E 3 [4]	74.96	70.81	78.72	78.50	79.50	79.83	80.82	78.82	75.82	76.82	73.39	67.27	73.45	67.20	72.01	71.34	63.59	60.72	89.66	86.67	66.83	54.83	72.93	60.99
7	FLUX.1 [dev] [37]	71.09	71.78	83.12	78.65	87.05	83.17	87.25	80.39	75.01	72.39	65.79	68.54	67.07	73.69	73.84	73.34	69.09	71.59	66.67	66.67	43.83	52.83	70.72	71.47
8	FLUX.1 [Pro] [37]	67.32	69.89	79.08	78.91	78.83	81.33	82.82	83.82	75.57	71.57	61.10	65.37	62.32	65.57	69.84	71.47	65.96	67.72	63.00	63.00	35.83	55.83	71.80	68.80
9	Midjourney V7 [52]	68.74	65.69	77.41	76.00	77.58	81.83	82.07	76.82	72.57	69.32	64.66	60.53	67.20	62.70	81.22	71.59	60.72	64.59	83.33	80.00	24.83	20.83	68.83	63.61
10	SD 3 [18]	67.46	66.09	78.32	77.75	83.33	79.83	82.07	78.82	71.07	74.07	61.46	59.56	61.07	64.07	68.84	70.34	50.96	57.84	66.67	76.67	59.83	20.83	63.23	67.34
11	SANA 1.5 [81]	67.15	65.73	79.66	77.08	79.83	77.83	85.57	83.57	73.57	69.82	61.50	60.67	65.32	56.57	69.96	73.09	62.96	65.84	80.00	80.00	17.83	15.83	71.07	68.83
12	Janus-Pro-7B [14]	66.50	65.02	79.33	78.25	79.33	82.33	78.32	73.32	80.32	79.07	59.71	58.82	66.07	56.20	70.46	70.84	67.22	59.97	60.00	70.00	28.83	33.83	65.84	60.25
13	Infinity [28]	62.07	62.32	73.08	75.41	74.33	76.83	72.82	77.57	72.07	71.82	56.64	54.98	60.44	55.57	74.22	64.71	60.22	59.71	80.00	73.33	10.83	23.83	54.28	56.89
14	PixArt-Σ [12]	62.00	58.12	70.66	75.25	69.33	78.83	75.07	77.32	67.57	69.57	57.65	49.50	65.20	56.57	69.96	61.72	66.59	54.59	83.33	70.00	1.83	1.83	62.11	52.41
15	Show-o [82]	59.72	58.86	73.08	75.83	74.83	79.83	78.82	78.32	65.57	69.32	53.67	50.38	60.95	56.82	68.59	68.59	66.46	56.22	63.33	66.67	3.83	2.83	55.02	50.92
16	LightGen [79]	53.22	43.41	66.58	47.91	55.83	47.33	74.82	45.82	69.07	50.57	46.74	41.53	62.44	40.82	61.71	50.47	50.34	45.34	53.33	53.33	0.00	6.83	50.92	50.55
17	Hunyuan-DiT [41]	51.38	53.28	69.33	69.00	65.83	69.83	78.07	73.82	64.07	63.32	42.62	45.45	50.20	41.57	59.22	61.84	47.84	51.09	56.67	73.33	0.00	0.83	40.10	44.20
18	Lumina-Next [99]	50.93	52.46	64.58	66.08	56.83	59.33	67.57	71.82	69.32	67.07	44.75	45.63	51.44	43.20	51.09	59.72	44.72	54.46	70.00	66.67	0.00	0.83	47.56	49.05

Table 11 | Quantitative results on PRISM-Bench [19] evaluated by Qwen2.5-VL-72B [2].

Rank	Model	Imagination			Entity			Text rendering			Style			Affection			Composition			Long text			Overall↑		
		Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.
1	GPT-Image-1 [High] [55]	89.1	53.3	66.6	87.3	81.0	84.1	66.7	86.8	76.8	87.3	87.8	87.5	88.1	79.8	84.0	92.2	84.9	88.5	77.2	77.5	77.4	82.7	78.7	80.7
2	Gemini 2.5-Flash-Image [25]	84.7	38.1	61.4	86.0	76.7	81.3	72.8	84.3	78.5	89.5	87.8	88.6	94.3	74.8	84.5	91.2	88.2	89.7	76.3	80.6	78.4	85.0	75.8	80.4
3	**Z-Image-Turbo**	65.7	50.1	57.9	75.7	82.3	79.0	59.6	84.9	72.2	76.7	88.2	82.4	85.1	87.4	86.2	89.0	90.2	89.6	69.8	79.0	74.4	74.5	80.3	77.4
4	Seedream 3.0 [21]	75.8	38.0	56.9	81.3	74.2	77.7	58.8	74.0	66.4	84.4	84.1	84.2	90.5	74.6	82.5	93.6	85.1	89.3	76.2	76.4	76.3	80.1	72.3	76.2
5	**Z-Image**	68.0	47.3	57.6	75.0	74.4	74.7	59.3	81.6	70.4	78.0	89.0	83.5	84.3	80.1	82.2	89.1	85.1	87.1	70.6	76.6	73.6	74.9	76.2	75.6
6	Qwen-Image [76]	75.5	37.4	56.5	79.5	64.5	72.0	57.9	71.2	64.5	86.6	84.4	85.5	89.9	70.4	80.1	93.9	79.5	86.7	76.8	70.9	73.8	80.0	68.3	74.1
7	FLUX.1-Krea-dev [37]	69.6	43.1	56.3	72.2	70.7	71.4	51.7	76.1	63.9	80.0	86.6	83.3	82.6	78.7	80.6	90.8	87.1	88.9	73.6	73.4	73.5	74.4	73.7	74.0
8	HiDream-I1-Full [7]	73.0	44.0	58.5	76.3	72.8	74.5	60.5	76.4	68.4	81.4	81.5	81.4	90.0	76.6	83.3	88.5	80.3	84.4	66.3	48.6	57.4	76.6	68.6	72.6
9	SD3.5-Large [18]	66.7	43.4	55.0	76.8	72.7	74.8	53.6	73.1	63.3	77.3	78.2	77.7	85.6	73.9	79.7	87.8	80.9	84.3	65.8	52.2	59.0	73.4	67.8	70.6
10	HiDream-I1-Dev [7]	68.8	45.8	57.3	73.5	68.1	70.8	56.7	75.7	66.2	70.2	77.4	73.8	88.2	74.3	81.2	84.7	78.5	81.6	64.0	49.3	56.6	72.3	67.0	69.6
11	FLUX.1-dev [37]	65.5	42.9	54.2	70.6	61.9	66.2	52.3	73.0	62.6	72.6	74.2	73.4	86.0	72.9	79.4	87.4	75.8	81.6	70.5	53.8	62.1	72.1	64.9	68.5
12	SD3.5-Medium [18]	65.1	34.7	49.9	72.5	70.9	71.7	36.6	64.5	50.5	75.5	80.0	77.7	81.8	73.9	77.9	85.4	81.0	83.2	63.5	50.6	57.0	68.6	65.1	66.8
13	SD3-Medium [18]	64.3	37.7	51.0	69.4	63.3	66.3	38.5	63.3	50.9	74.6	79.5	77.0	80.5	75.5	78.0	85.6	79.5	82.5	63.4	50.3	56.8	68.0	64.2	66.1
14	FLUX.1-schnell [37]	62.8	35.6	49.2	64.8	56.8	60.8	54.3	68.1	61.2	70.3	71.5	70.9	75.4	65.9	70.6	81.7	75.6	78.6	68.7	54.4	61.5	68.3	61.1	64.7
15	Janus-Pro-7B [14]	65.0	38.8	51.9	68.6	63.5	66.0	23.1	50.3	36.7	75.2	72.9	74.0	80.7	68.0	74.3	82.4	71.1	76.7	63.9	49.0	56.4	64.9	59.4	62.1
16	Bagel [15]	68.0	45.0	56.5	67.6	53.4	60.5	29.4	42.3	35.8	69.0	69.7	69.3	87.1	66.7	76.9	86.6	69.2	77.9	64.5	50.2	57.3	67.5	56.6	62.0
17	Bagel-CoT [15]	68.0	44.1	56.0	67.6	53.4	60.5	29.4	42.3	35.8	69.0	69.7	69.3	87.1	66.7	76.9	86.6	69.2	77.9	64.5	50.2	57.3	67.5	56.6	62.0
18	Playground [39]	59.0	39.0	49.0	69.4	56.7	63.0	15.3	31.9	23.6	74.6	74.6	74.6	88.8	66.0	77.4	72.2	61.3	66.7	56.0	35.3	45.6	62.2	52.1	57.1
19	SDXL [57]	54.5	34.1	44.3	71.1	65.0	68.0	18.6	37.3	27.9	71.7	72.6	72.1	78.7	66.5	72.6	72.2	67.8	70.0	54.1	34.5	44.3	60.1	54.0	57.0
20	SD2.1 [63]	48.9	28.4	38.6	66.0	57.6	61.8	16.7	31.4	24.0	62.7	66.5	64.6	68.5	62.1	65.3	64.8	58.3	61.5	50.7	29.8	40.2	54.0	47.7	50.8
21	SD1.5 [63]	40.7	23.7	32.2	61.2	52.7	56.9	11.4	24.1	17.8	56.7	61.5	59.1	66.9	60.7	63.8	57.5	53.4	55.4	47.3	26.8	37.0	48.8	43.3	46.0

5.2.2. Instruction-based Image Editing

ImgEdit. Table 13 shows the evaluation of Z-Image-Edit on the ImgEdit Benchmark [87], where the metric combines instruction completion and visual quality. Across 9 common editing tasks, Z-Image-Edit

Table 12 | Quantitative results on PRISM-Bench-ZH [19] evaluated by Qwen2.5-VL-72B [2].

Rank	Model	Imagination			Entity			Text rendering			Style			Affection			Composition			Long text			Overall↑		
		Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.	Ali.	Aes.	Avg.
1	GPT-Image-1 [High] [55]	**73.0**	**37.6**	**55.3**	80.4	82.1	**81.3**	73.1	89.9	81.5	**77.1**	**92.4**	**84.8**	78.0	77.8	**77.9**	**91.9**	85.7	**88.8**	72.4	**76.3**	**74.4**	**78.0**	**77.4**	**77.7**
2	**Z-Image**	69.5	34.1	51.6	70.6	73.7	72.2	**76.8**	**90.0**	**83.4**	74.1	88.2	81.2	77.6	73.5	75.5	89.3	**88.0**	88.6	71.6	75.6	73.6	75.7	74.9	75.3
3	**Z-Image-Turbo**	64.1	37.2	50.7	72.9	**82.4**	77.6	69.4	89.7	79.6	72.9	89.2	81.0	74.0	**80.9**	77.5	87.2	85.8	86.5	71.7	74.8	73.3	73.1	77.1	75.1
4	Seedream 3.0 [21]	71.4	36.6	54.0	74.8	73.8	74.3	70.7	88.0	79.4	74.1	88.0	81.1	**79.0**	71.4	75.2	90.3	83.2	86.8	**73.0**	71.2	72.1	76.2	73.2	74.7
5	Qwen-Image [76]	71.4	29.9	50.7	74.7	67.8	71.3	64.3	73.1	68.7	75.2	83.2	79.2	77.3	64.5	70.9	89.8	74.1	82.0	72.6	65.8	69.2	75.0	65.5	70.3
6	Bagel-CoT [15]	64.4	36.6	50.5	62.6	53.8	58.2	25.2	51.9	38.6	65.4	76.7	71.1	74.0	65.0	69.5	81.3	71.3	76.3	61.4	46.6	54.0	62.0	57.4	59.7
7	Bagel [15]	64.6	36.3	50.5	62.7	55.5	59.1	18.6	26.3	22.5	66.0	76.6	71.3	74.9	66.2	70.6	81.3	72.2	76.8	62.4	47.3	54.9	61.5	54.3	57.9
8	HiDream-I1-Full [7]	51.2	30.8	41.0	60.1	61.3	60.7	20.7	40.6	30.7	64.5	73.8	69.2	65.2	69.1	67.2	72.4	69.0	70.7	57.1	42.8	50.0	55.9	55.3	55.6
9	HiDream-I1-Dev [7]	48.3	24.6	36.5	52.6	54.1	53.4	18.6	35.3	27.0	59.0	68.3	63.7	65.9	62.3	64.1	66.5	64.6	65.6	54.2	38.6	46.4	52.2	49.7	50.9

Table 13 | Quantitative Evaluation results on ImgEdit [87].

Rank	Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall↑
1	UniWorld-V2 [43]	4.29	**4.44**	**4.32**	**4.69**	**4.72**	4.41	4.91	3.83	4.83	**4.49**
2	Qwen-Image-Edit [2509] [76]	4.32	4.36	4.04	4.64	4.52	4.37	4.84	3.39	4.71	4.35
3	**Z-Image-Edit**	4.40	4.14	4.30	4.57	4.13	4.14	4.85	3.63	4.50	4.30
4	Qwen-Image-Edit [76]	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
5	GPT-Image-1 [High] [55]	**4.61**	4.33	2.90	4.35	3.66	**4.57**	**4.93**	**3.96**	**4.89**	4.20
6	FLUX.1 Kontext [Pro] [37]	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
7	OmniGen2 [78]	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
8	UniWorld-V1 [44]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
9	BAGEL [15]	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
10	Step1X-Edit [47]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
11	ICEdit [94]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
12	OmniGen [80]	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
13	UltraEdit [95]	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
14	AnyEdit [90]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
15	MagicBrush [92]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
16	Instruct-Pix2Pix [5]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88

Table 14 | Quantitative Evaluation results on GEdit-Bench [47].

Rank	Model	GEdit-Bench-EN			GEdit-Bench-CN		
		G_SC	G_PQ	G_O↑	G_SC	G_PQ	G_O↑
1	UniWorld-V2 [43]	**8.39**	**8.02**	**7.83**	-	-	-
2	Qwen-Image-Edit [2509] [76]	8.15	7.86	7.54	**8.08**	**7.89**	7.54
3	**Z-Image-Edit**	8.11	7.72	7.57	8.03	7.80	**7.54**
4	Qwen-Image-Edit [76]	8.00	7.86	7.56	7.82	7.79	7.52
5	GPT-Image-1 [High] [55]	7.85	7.62	7.53	7.67	7.56	7.30
6	Step1X-Edit [47]	7.66	7.35	6.97	7.20	6.87	6.86
7	BAGEL [15]	7.36	6.83	6.52	7.34	6.85	6.50
8	OmniGen2 [78]	7.16	6.77	6.41	-	-	-
9	FLUX.1 Kontext [Pro] [37]	7.02	7.60	6.56	1.11	7.36	1.23
10	FLUX.1 Kontext [Dev] [37]	6.52	7.38	6.00	-	-	-
11	OmniGen [80]	5.96	5.89	5.06	-	-	-
12	UniWorld-V1 [44]	4.93	7.43	4.85	-	-	-
13	MagicBrush [92]	4.68	5.66	4.52	-	-	-
14	Instruct-Pix2Pix [5]	3.58	5.49	3.68	-	-	-
15	AnyEdit [90]	3.18	5.82	3.21	-	-	-

shows competitive editing performance with leading models , especially object addition and extraction.

GEdit. We also evaluate Z-Image-Edit on the GEdit-Bench [47], which evaluates visual naturalness (G_PQ) and bilingual instruction following (G_SC). GEdit-Bench-EN abd GEdit-Bench-CN adopt English and Chinese instructions in the evaluation, respectively. As shown in Table 14, Z-Image-Edit achieves 3rd rank, demonstrating robust bilingual editing capabilities.

5.3. Qualitative Evaluation

To further demonstrate the visual generation capacity of Z-Image [4], we first give the qualitative comparison against state-of-the-art open-source models (Lumina-Image 2.0 [58], Qwen-Image [76], Hun-

---

[4]In the section, all results of Z-Image are generated by our Turbo version.

yuanImage 3.0 [8], and FLUX 2 dev [36]) and close-source models (Imagen 4 Ultra [26], Seedream 4.0 [64] and Nano Banana Pro [27]). We then show the editing capacity of our Z-Image-Edit. We next show the examples of how reasoning capacity and world knowledge are injected by our prompt enhancer. We finally show that the emerging multi-lingual and multi-cultural understanding capacity of our Z-Image.

5.3.1. Superior Photorealistic Generation

As shown in Figure 16 and 17, Z-Image-Turbo shows excellent character close-up generation (*e.g.*, the skin details on a man's face and a girl's tears). When asked to generate multi-expression portraits of one person (Figure 18), Z-Image-Turbo can produce images that are more aesthetically pleasing and have more realistic expressions, while Qwen-Image, HunyuanImage3.0, FLUX 2 dev, and Seedream 4.0 would sometimes generate exaggerated and unrealistic expressions, thus lacking authenticity and beauty. Moreover, when generating a scene captured by a mobile phone (Figure 19 and 20), Z-Image-Turbo shows strong performance in the authenticity of both the person and the background, as well as the aesthetic appeal of layout and posture. while Qwen-Image, HunyuanImage3.0, and FLUX 2 dev would generate unrealistic things (*e.g.*, clothes that remain completely unsoaked in the heavy rain).

5.3.2. Outstanding Bilingual Text Rendering

Figure 21 and Figure 22 show the qualitative comparison of Chinese and English text rendering. As shown in Figure 21 and Figure 22, Z-Image-Turbo accurately rendered the required text while maintaining the aesthetic appeal and authenticity of other parts (*e.g.*, the authenticity of the human face in Figure 21 and the layout of the scene in Figure 22). Note that this is comparable to the leading closed-source model Nano Banana Pro, and surpasses other candidates. When rendering text in poster design (Figure 23 and Figure 24), Z-Image-Turbo not only presents correct text rendering, but also designs a more aesthetically pleasing and realistic poster. For example, as shown in Figure 24), Qwen-Image, HunyuanImage3.0, FLUX 2 dev, and Imagen 4 Ultra make errors when rendering very small characters, Seedream4.0 and Nano Banana Pro make errors of repeatedly rendering the text, while Z-Image-Turbo gets the poster with the right rendered text and satisfactory design.

5.3.3. Instruction-following Editing

The first two columns of Figure 25 demonstrates the ability of Z-Image-Edit to handle complex composite prompts. For example, the top simultaneously switches the background to the Sydney Opera House, inserting a specific object (a sign reading "Z-Image"), and removing the character's backpack. The bottom row also illustrates precise control over multiple subjects. The last two columns of Figure 25 also illustrates that Z-Image-Edit can accurately modify textual content according to a bounding-box–based location constraint (left) and keep characters consistent when transforming the image (right).

5.3.4. Enhanced Reasoning Capacity and World Knowledge through Prompt Enhancer

As demonstrated in Figure 15 and Figures 26-27, our prompt enhancer leverages a structured reasoning chain – comprising core subject analysis, problem solving/world knowledge injection, aesthetic enhancement, and comprehensive description – to equip the model with logical reasoning and world knowledge capabilities. This allows the model to handle diverse tasks, ranging from solving complex logical puzzles (*e.g.*, the chicken-and-rabbit problem) and interpreting user intent (*e.g.*, visualizing classical poetry or inferring scenes from coordinates) to performing text rendering and question answering.

In the context of image editing, prompt enhancer is also crucial for addressing ambiguous or unclear intentions, as well as for injecting world knowledge and enabling reasoning, similar to how it functions in text-to-image generation, as shown in Figure 28-29. For example, in Figure 29, the wrong dish is made because of a lack of reasoning about the relationship between the ingredients and the dish, while prompt enhancer can make up for this.

5.3.5. Emerging Multi-lingual and Multi-cultural Understanding Capacity

After trained with bilingual data, we are surprised to find that Z-Image has initially emerged with the ability to handle multilingual input. As shown in Figure 30, Z-Image can not only understand prompts in multiple languages but also generate images that align with local cultures and landmarks.

Input prompt 女生大哭。(The girl is crying.)
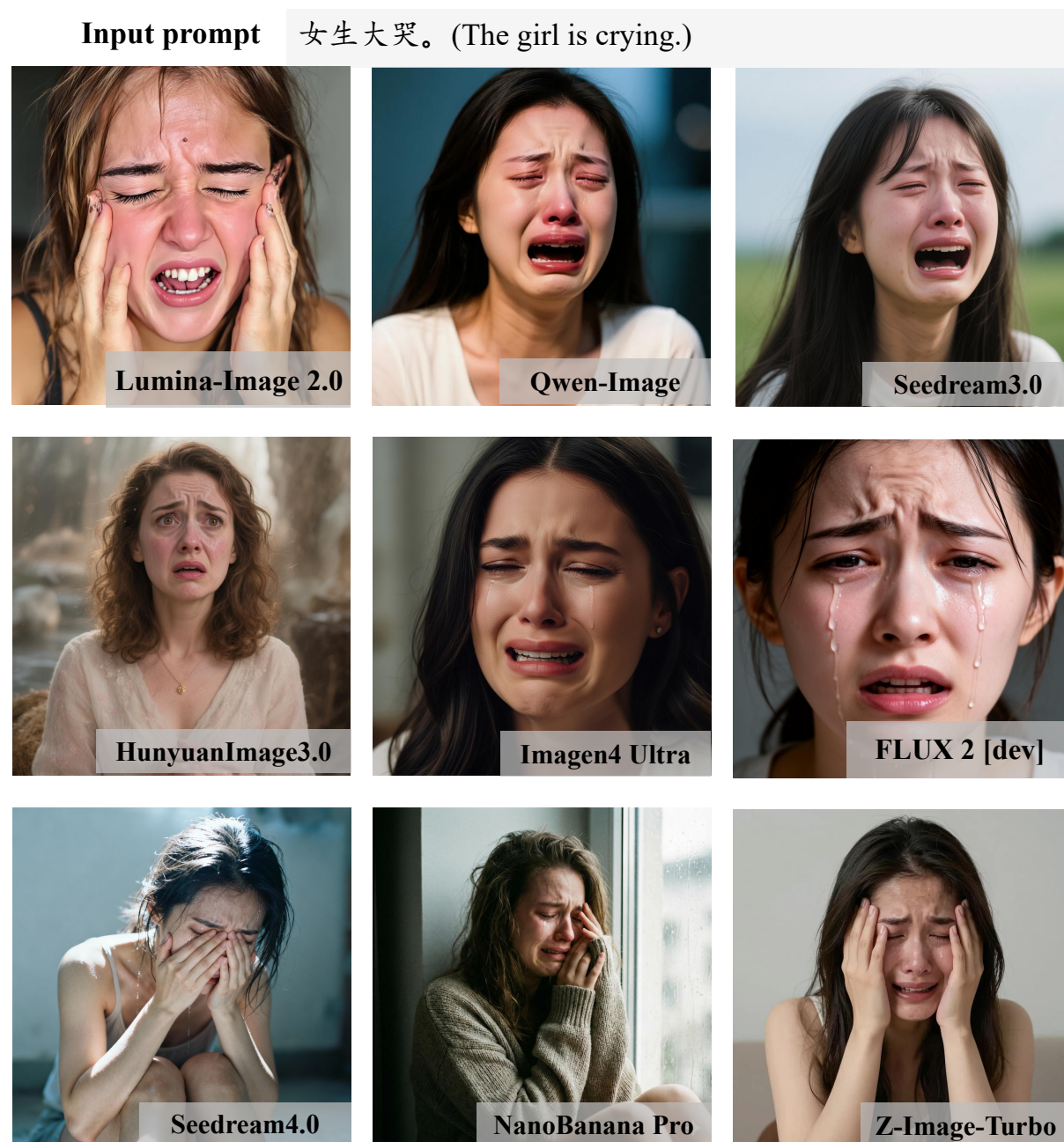
Figure 16 | Comparison of close-up portrait generation, which indicates that Z-Image exhibits strong capabilities in character emotion and skin texture rendering. Better to zoom in to check the subtle expressions and the texture of the skin.

Input prompt 男人沉思。(The man is lost in thought.)

Figure 17 | Comparison of close-up portrait generation, which indicates that Z-Image exhibits strong capabilities in character emotion and skin texture rendering. Better to zoom in to check the subtle expressions and the texture of the skin.

一个三乘三的网格拼贴画，在纯白色的背景下，展示了同一位年轻亚洲女性的九幅工作室肖像，她有着黑色的长发和细细的刘海。这些照片捕捉了各种各样有趣而感性的面部表情。从最上面一排开始，从左到右：她用右眼眨眼，同时调皮地伸出舌头。她热情地笑着，眼睛闭得紧紧的，一种害羞或高兴的样子。她双手捂着嘴，睁大眼睛直视镜头，似乎是惊讶或害羞，脸颊明显泛红。中间一排：她直视镜头，表情中性，微微撅着嘴。她笑到一半，头略微向后仰，眼睛眯着，嘴巴张得大大的，露出了牙齿。她用左眼眨了眨眼睛，同时微微皱了皱眉。最下面一排：她左手举着和平手势（v字），眼睛睁大大的，舌头微微伸出。她的两只手垂直地放在鼻子和眼睛旁边，嘴巴张开成一个圆形的"O"形，眼睛闭着，好像在喊叫或偷看。她用手托着左脸颊，带着一种温柔、沉思的表情，微微离开镜头向右看。所有镜头的灯光都是明亮而均匀的。

Figure 18 | Comparison of complex close-up portrait generation, which indicates that Z-Image-Turbo has a strong ability in rendering character expressions and skin textures, as well as generating aesthetic images. Better to zoom in to check the subtle expressions.

33

Input prompt	一位穿着白衬衫的女人在雨夜独自走在街头。她的神情中透露出失落，微微低着头，雨丝轻轻落下，在湿漉漉的路面上形成了小水洼。街灯洒下柔和的光芒，映照出她头发和衣服上的雨滴。背景中模糊的城市建筑增添了忧郁孤独的氛围。(A woman dressed in a white shirt walked alone along the street on a rainy night. Her expression conveyed a sense of loss as she slightly lowered her head. Raindrops fell gently, forming small puddles on the wet pavement. The streetlights emitted a soft glow, reflecting off the raindrops that clung to her hair and clothing. In the background, blurred city buildings contributed to an atmosphere of melancholy and solitude.
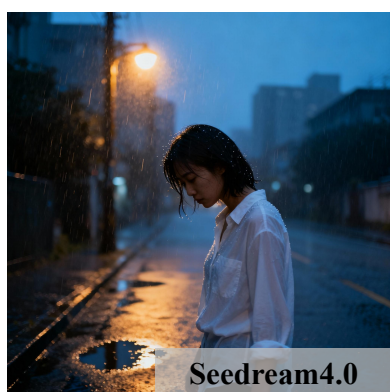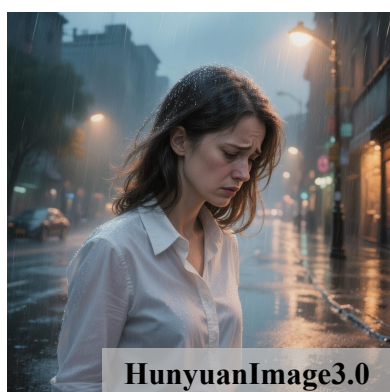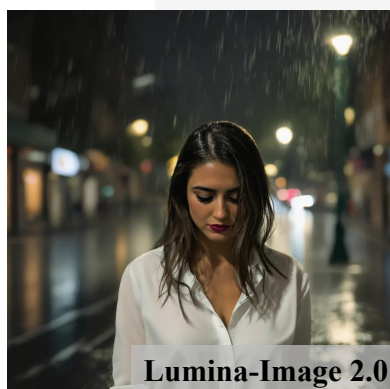
Figure 19 | Comparison of scene shooting, which indicates that Z-Image-Turbo shows strong performance in the authenticity of both the person and the background, as well as the aesthetic appeal of layout and posture. Better to zoom in to check the texture of the clothes and hair.

	An Asian chef is stir-frying at a roadside stall in an old-fashioned black iron pot,
Input prompt	creating a vibrant and lively atmosphere. (一位亚洲厨师正在路边小摊上用老式的黑铁锅炒菜，营造出充满活力和活力的氛围。)

Figure 20 | Comparison of scene shooting, which indicates that Z-Image-Turbo shows strong performance in the authenticity of both the person and the background, as well as the aesthetic appeal of layout and posture. Better to zoom in to check the details.

Input prompt	镜头前，一位身着印有 "Z-Image：让创意即刻轻装上阵" 白 T 恤的中国女教师笑意盈盈，手持白色粉笔，她身后的黑板上，写着："Z-Image模型亮点：1. 文字艺术家：中英双语秒级渲染，排版自动大片级。2. 真实感魔术师：摄影级场景布局，超真实质感，一眼告别 AI 味。3. 速度狂魔：创新S3-DiT架构，8 步推理，6B 参数，快到让你怀疑GPU 开了挂。"

Figure 21 | Comparison of complex Chinese text rendering. It shows that only Z-Image-Turbo and Nano Banana Pro can accurately generates the expected Chinese couplet. Better to zoom in to check the correctness of the rendered text and the authenticity of the person.

Input prompt	A contemporary loft interior with feature wall displaying large mixed-font typography "Every moment presents an opportunity to create something meaningful, to inspire others, to leave a legacy that transcends time and touches the hearts of those around us." against geometric patterns, industrial aesthetic, natural window lighting.
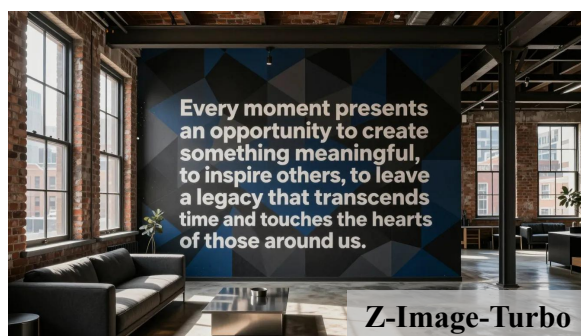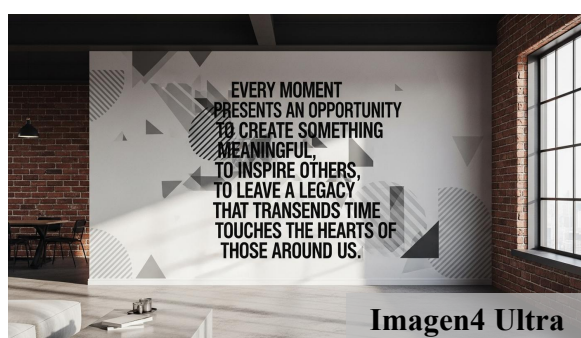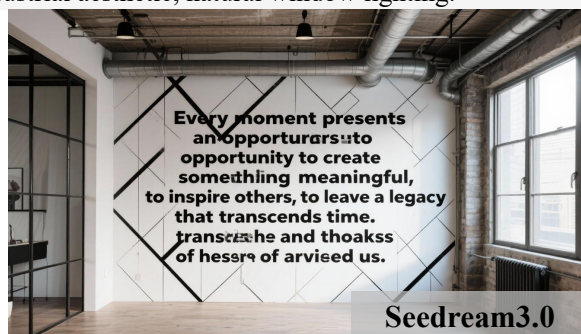
Figure 22 | Comparison of complex English text rendering. It shows that only Z-Image-Turbo and Nano Banana Pro can accurately generates the expected English couplet. Better to zoom in to check the correctness of the rendered text and the layout of the scene.

设计一张名为"造相美食"的烘焙店商业广告海报，以双色调背景（上部为浅桃色，下部为灰玫瑰色）展示了三款不同的水果蛋糕。在上半部分，用红棕色大号中文字体写着"轻甜&轻负担"。其下方是较小的白色文字"（当季新鲜水果）"。主标题下方印有红棕色无衬线字体的英文"Z-Image Cake"。左侧，一个装饰着红色覆盆子、石榴籽和绿色嫩枝的蛋糕放在一个浅色木质底座。右侧，另一个放在更高木质底座上的蛋糕顶部有无花果、葡萄和蓝莓。右下方，第三个蛋糕放在一个白色盘子上，覆盖着橙色柿子酱，顶部有柿子和一朵干花。在蛋糕之间，一杯石榴籽洒落在台面上。左下角有两个新鲜的柿子。海报为每个蛋糕都附有描述性文字："留红石 ¥119"，配料为"酸奶芝士奶油"，规格为"规格：10.5cm*10.5cm"；"有花香 ¥139"，配料为"酸奶芝士奶油·无花果酱"，规格为"规格：10.5cm*10.5cm"；以及"秋柿早 ¥129"，配料为"柿子奶油·柿子酱"，规格为"规格：10.5cm*10.5cm"。品牌名"造相美食"位于左下角。右下角写着"NEW TASTE"和"（秋冬尝鲜）"。该图像是一张高质量的影棚摄影作品，光线柔和温暖，营造出诱人而洁净的美感。

Figure 23 | Comparison of Chinese text rendering in poster design. Z-Image-Turbo not only presents correct text rendering, but also designs a more aesthetically pleasing and realistic poster. Better to zoom in to check the correctness of the rendered text and the fidelity of the food.

Figure 24 | Comparison of English text rendering in poster design. Only Z-Image-Turbo presents correct text rendering with a pleasing and realistic poster. Better to zoom in to check the correctness of the rendered text and the details of the poster.

背景切换为悉尼歌剧院，手上拿着一个牌子写着"Z-Image"，删除背包。

The background changes to the Sydney Opera House, and the person is holding a sign that reads "Z-Image" while removing his backpack.

左边的衣服改成蓝色，中间的衣服改成绿色，右边的衣服改成黄色，三个人的表情改为中性，同时删除所有发卡。

Change the clothes on the left to blue, the clothes in the middle to green, and the clothes on the right to yellow. Change the expressions of all three people to neutral and delete all hair clips.

红色框里的文字修改为"由地"。天空改为更晴朗的蓝天白云，阳光灿烂。

The text in the red box should be changed to "由地". Change it to a clearer picture of blue sky, white clouds, and bright sunshine.

她现在蹲在草地上，轻轻地抚摸着一只哈士奇犬，背景是开阔的户外环境和柔和的天空，同时保持着同样的衣着、身份、个性、种族、性别和面容。

She is now crouching on a grassy field, gently petting a husky dog, with an open outdoor environment and a soft sky in the background, while keeping the same clothing, identity, personality, race, gender, and face.

Figure 25 | **The first two columns**: Mixed-instruction editing across various tasks in Z-Image-Edit. **The last two columns**: Text editing (with bounding box) and identity-preservation editing in Z-Image-Edit.
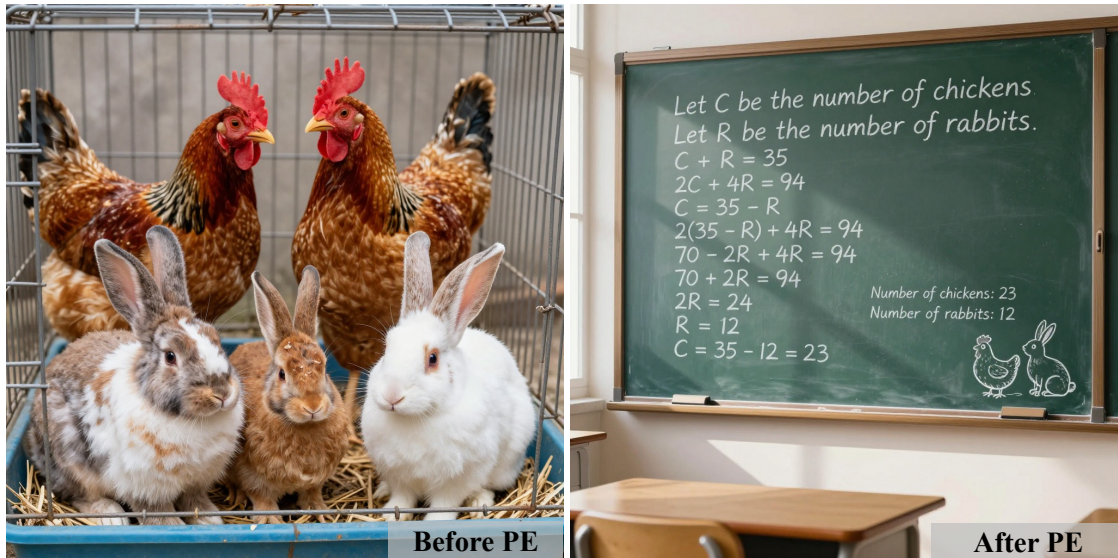
Input: Given that chickens and rabbits are in the same cage, there are a total of 35 heads and 94 feet. Find the number of chickens and rabbits.

Summarized Reasoning:
1. Analysis & Language Locking: Identified the user input as a mathematical word problem ("chickens and rabbits") in English, determining that the output prompt must also be in English. Recognized the need for "Generative Reasoning" to visualize the solution rather than just the problem.

2. Solution Logic (Generative Reasoning) : Solved the math problem step-by-step using a system of linear equations to find the specific values (23 chickens, 12 rabbits). Decided to visualize this process as a "blackboard derivation" in a classroom setting, which serves as the logical visual carrier.

3. Aesthetic & Scene Construction: Designed a realistic classroom atmosphere with natural lighting, specific textures (chalk dust, wood grain), and a shallow depth of field to focus on the content.

4. Text Precision: Transcribed the exact mathematical steps and final answer (e.g., "2C + 4R = 94", "R = 12") into explicit text instructions, ensuring all content was enclosed in double quotes for the image generation model.

Output: A classroom setting with a large green chalkboard on a wooden frame, illuminated by soft morning light from a window on the left. On the chalkboard, written in clear white chalk handwriting: "Let C be the number of chickens", "Let R be the number of rabbits", "C + R = 35", "2C + 4R = 94", "C = 35 - R", "2(35 - R) + 4R = 94", "70 - 2R + 4R = 94", "70 + 2R = 94", "2R = 24", "R = 12", "C = 35 - 12 = 23", "Number of chickens: 23", "Number of rabbits: 12". A small hand-drawn sketch of a chicken and a rabbit appears in the bottom right corner. Foreground features an empty oak desk and chair, slightly out of focus. Dust particles float in the light beam. Textured chalkboard surface shows subtle chalk dust residue. Wooden frame has visible grain. Color scheme: muted greens, warm browns, and crisp white. Shallow depth of field emphasizes the chalkboard.

Figure 26 | Showcases of prompt enhancer for logical reasoning.

Input: 帮我给《登科后》配图，最出名的两句

Summarized Reasoning:

1. 语言与核心锁定（Language & Core Locking）：
   确认输入为中文，锁定输出必须为中文。
   明确核心题材为唐诗《登科后》，并锁定最著名的两句诗文内容（"春风得意马蹄疾，一日看尽长安花"）。

2. 生成式推理与画面转化（Generative Reasoning）：
   将抽象的诗意（得意、喜悦）转化为具象的视觉场景：一位身着红袍的新科进士，在春日阳光下骑马疾驰穿过长安街道，周围花团锦簇。

3. 美学与细节设定（Aesthetic & Detail Injection）：
   设定氛围为明亮、欢快的春日午后。
   细化元素包括唐代木质建筑、丝绸服饰质感、粉白花朵与飞扬的尘土，构建富有层次的空间感。

4. 文字元素的精确植入（Text Transcription）：
   为了忠实还原诗意，决定将诗句作为视觉元素（书法卷轴）置于画面底部。
   严格遵守指令，将诗句内容一字不差地转录并使用英文双引号包裹（""春风得意马蹄疾，一日看尽长安花。""）。

Output:
一位年轻男子身着唐代红色丝绸官袍，骑在一匹棕色骏马上，马蹄轻快踏过长安城街道石板路，男子面容欣喜，目光前视，春风拂动其衣袖与马鬃。背景为古代长安街景，两旁排列粉白色盛开的桃树，花瓣随风飘落，远处可见青瓦木构建筑、城楼与商铺旗帜，街边有数名唐代服饰的民众驻足观看。春日午后阳光明亮，柔和阴影投在地面，整体色彩以暖色调为主：蓝天、嫩绿柳枝、红袍与棕马形成对比，材质细节包括丝绸光泽、马匹毛发纹理、木质建筑粗糙表面及花瓣柔软质感。画面底部中央，一个展开的米白色宣纸卷轴，宽度占画面三分之一，上面以黑色毛笔书法字体清晰写着："春风得意马蹄疾，一日看尽长安花。"卷轴边缘有竹制轴杆，固定于地面。构图采用中景视角，焦点在骑手与马匹，空间层次分明，前景为飘落花瓣，中景街道，背景建筑渐隐。

Figure 27 | Showcases of prompt enhancer for world knowledge injection. Given the poem title "After Passing the Imperial Examination" (《登科后》), the baseline (Left) lacks cultural context. Our method (Right) leverages LLM priors to retrieve specific historical details (e.g., the galloping horse, red official robe) and the famous couplet: "春风得意马蹄疾，一日看尽长安花。", the reasoning module (center) translates these literary semantics into visual cues, ensuring a culturally faithful rendering with precise text transcription.

| Input Image | Output Image (Before PE) | Output Image (After PE) |
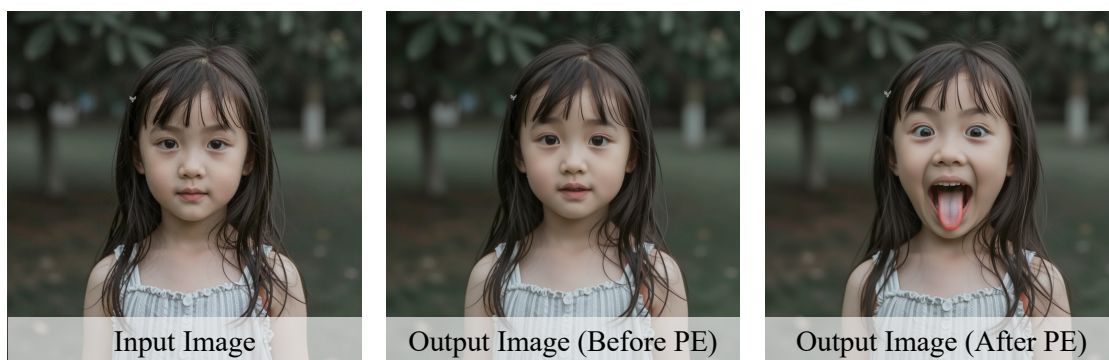
Input: 表情包 （Emoticons）

Reasoning:
1. Source Analysis: A little girl stands in an outdoor park, with blurred trees and green vegetation in the background. She has long, black hair with bangs neatly covering her forehead, and a silver butterfly hair clip on the left side of her hair. She wears a light-colored striped sundress with small button decorations. Her expression is calm, her eyes look directly at the camera, her facial features are clear, and her skin is smooth. The overall image is realistic, using natural light that is soft and even, creating a tranquil and refreshing atmosphere. The composition is centrally symmetrical, focusing on the little girl's face, with a blurred background that emphasizes the subject. The image has a cool color tone, mainly gray-green and light blue, creating a quiet visual experience.

2. Edit Plan: Preserve the little girl's overall appearance, hairstyle, clothing, and background environment, maintaining the original natural lighting and composition. Modify the little girl's expression to an exaggerated, humorous one, such as opening her mouth wide, staring, or making a comical tongue-out gesture, to conform to common emoji styles. Ensure the new expression seamlessly integrates with the original character features, without altering other details such as hair accessories, clothing textures, and the background.

Output: The little girl's facial expressions transform into exaggerated, comical gestures, such as a wide-open mouth, bulging eyes, or sticking out her tongue, conveying a humorous and playful mood. Despite the change in expression, her original long hair, butterfly hair clip, striped sundress, and the trees and greenery in the background are retained. The overall style shifts from the original realistic tranquility to a comedic, meme-like style, but the character's posture, clothing details, and background environment remain unchanged. The image continues to use natural lighting and maintains a cool color palette, but the dynamic changes in facial expression become the visual focus, highlighting its fun and viral potential as a meme.

Figure 28 | Showcases of prompt enhancer in image editing for handling ambiguous and unclear instructions.

| Input Image | Output Image (Before PE) | Output Image (After PE) |

Input: 做一道菜 （Make a dish）

Reasoning:
1. Source Analysis: The kitchen countertop is laden with ingredients, including a cracked egg, the yolk still intact in half the shell, the other shell lying empty beside it. Next to it is a bright red, plump tomato, glistening with moisture, its green stem clearly visible. Several scallions are neatly arranged in the background, their stems vibrant green and leaves slender. The countertop is made of wood with a warm brown hue. In the background, a stainless steel stovetop, white cabinets, and the sink area are visible. The overall environment is clean and tidy, with bright, natural lighting. Sidelight from the window creates a soft glow on the food, creating a fresh, pre-cooking atmosphere.

2. Edit Plan: Remove the tomatoes, scallions, and eggs (including shells and yolks) from the original image, while preserving the kitchen background, wooden countertop, and overall lighting effects. Add a cooked dish, specifically scrambled eggs with tomatoes, to the original ingredients' locations, placed in the center of a white ceramic plate. The new dish should have a texture of golden-yellow egg pieces mixed with red tomato chunks, garnished with scallions, and slightly glossy to convey a steaming appearance.

Output: The original ingredients (tomatoes, scallions, and eggs) have been removed, leaving only the original kitchen background, the fragrant countertop, and the natural, pristine conditions. In the center of the foreground is a white ceramic plate containing steaming hot scrambled eggs with tomatoes. The tomato chunks are golden yellow, the diced tomatoes are red and juicy, and the scallions are bright green on the countertop. The surface is slightly glistening with oil, giving the impression that the dish was just cooked.

Figure 29 | Showcases of prompt enhancer in image editing for world knowledge injection and reasoning.

Figure 30 | Emerging Multi-lingual and Multi-cultural Understanding Capacity of Z-Image-Turbo. It shows that Z-Image-Turbo can not only understand prompts in multiple languages but also leverage its world knowledge to generate images that align with local cultures and landmarks.

6. Conclusion

In this report, we introduce the **Z-Image series**, a suite of high-performance 6B-parameter models built upon a Scalable Single-Stream Diffusion Transformer (S3-DiT). Challenging the prevailing "scale-at-all-costs" paradigm, we propose a holistic end-to-end solution anchored by four strategic pillars: (1) a curated, efficient data infrastructure; (2) a scalable single-stream architecture; (3) a streamlined training strategy; and (4) advanced optimization techniques for high-quality and efficient inference, encompassing PE- aware supervised fine-tuning, few-step distillation, and reward post-training.

This synergy allows us to complete the entire workflow **within 314K H800 GPU hours** at a total cost of **under $630K**, delivering top-tier photorealistic synthesis and bilingual text rendering. Beyond the robust base model, our pipeline yields **Z-Image-Turbo**, which enables **sub-second inference (<1s)** on an enterprise-grade H800 GPU and fits **comfortably within 16G VRAM consumer-grade hardware**. Additionally, we develop **Z-Image-Edit**, an editing model **efficiently derived** via our omni-pretraining paradigm. Through this pipeline, we provide the community with a blueprint for developing accessible, budget-friendly, yet state-of-the-art generative models.

7. Authors

7.1. Core Contributors[5]

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, Zhaohui Hou, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, Shilin Zhou

7.2. Contributors[6]

Chenglin Cai, Yujing Dou, Yan Gao, Minghao Guo, Songzhi Han, Wei Hu, Yuyan Huang, Xu Li, Zefu Li, Heng Lin, Linhong Luo, Qingqing Mao, Jingyuan Ni, Chuan Qin, Lin Qu, Jinghua Sun, Peng Wang, Ping Wang, Shanshan Wang, Xuecong Wang, Yi Wang, Yue Wang, Tingkun Wen, Junde Wu, Minggang Wu, Xiongwei Wu, Yi Xin, Haibo Xing, Xiaoxiao Xu, Ze Xu, Xunliang Yang, Shuting Yu, Yucheng Zhao, Jianan Zhang, Jianfeng Zhang, Jiawei Zhang, Qiang Zhang, Xudong Zhao, Yu Zheng, Haijian Zhou, Hanzhang Zhou

---

[5]Core Contributors are listed in alphabetical order of the last name.
[6]Contributors are listed in alphabetical order of the last name.

References

[1] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 929–947, 2024.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Qi Cai, Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Hidream-i1: An open-source high-efficient image generative foundation model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13636–13639, 2025.

[8] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

[9] Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*, 2025.

[10] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pages 386–402. Springer, 2024.

[11] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

[12] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[13] Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*.

[14] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[15] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining, 2025. *URL https://arxiv. org/abs/2505.14683*.

[16] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

[17] Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv preprint arXiv:2503.23461*, 2025.

[18] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[19] Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. Flux-reason-6m and prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*, 2025.

[20] Peng Gao, Le Zhuo, Chris Liu, , Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024.

[21] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

[22] Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

[23] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[24] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.

[25] Google. Gemini 2.5 flash & 2.5 flash image model card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf`, 2025.

[26] Google. Imagen 4 model card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Imagen-4-Model-Card.pdf`, 2025.

[27] Google. Nano banana pro. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf`, 2025.

[28] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15733–15744, 2025.

[29] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *Advances in Neural Information Processing Systems Workshops (NeurIPS Workshops)*, 2021.

[30] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

[31] Ideogram. Ideogram v3. `https://ideogram.ai`, 2025.

[32] Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Xin Jin, David Liu, Zhen Li, Mengmeng Wang, Peng Gao, and Harry Yang. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649*, 2025.

[33] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184, 2024.

[34] Kuaishou Kolors Team. Kolors 2.0. https://app.klingai.com/cn/, 2025.

[35] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2023.

[36] Black Forest Labs. FLUX.2: State-of-the-Art Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

[37] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

[38] LeaderGPU. Gpu server rental pricing. https://www.leadergpu.com/, 2025. Accessed: November 2025.

[39] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

[40] Yuhan Li, Xianfeng Tan, Wenxiang Shang, Yubo Wu, Jian Wang, Xuanhong Chen, Yi Zhang, Hangcheng Zhu, and Bingbing Ni. Ragdiffusion: Faithful cloth generation via external knowledge assimilation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17485–17495, 2025.

[41] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

[42] Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. Visualcloze: A universal image generation framework via visual in-context learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2025.

[43] Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, and Li Yuan. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025.

[44] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

[45] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[46] Dongyang Liu, David Liu, Peng Gao, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, and Steven Hoi. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. *arXiv preprint*, 2025.

[47] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

[48] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[49] Yiting Lu, Jiakang Yuan, Zhen Li, Shitian Zhao, Qi Qin, Xinyue Li, Le Zhuo, Licheng Wen, Dongyang Liu, Yuewen Cao, et al. Omnicaptioner: One captioner to rule them all. *arXiv preprint arXiv:2504.07089*, 2025.

[50] Chunjie Luo, Jianfeng Zhan, Xiaohe Xue, Lei Wang, Rui Ren, and Qiang Yang. Cosine normalization: Using cosine similarity instead of dot product in neural networks. In *International conference on artificial neural networks*, pages 382–391. Springer, 2018.

[51] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.

[52] Midjourney. Midjourney v7. https://www.midjourney.com/home, 2025.

[53] Quang-Huy Nguyen, Cuong Q Nguyen, Dung D Le, and Hieu H Pham. Enhancing few-shot image classification with cosine transformer. *IEEE Access*, 11:79659–79672, 2023.

[54] Hiroyuki Ootomo, Akira Naruse, Corey Nolet, Ray Wang, Tamas Feher, and Yong Wang. Cagra: Highly parallel graph construction and approximate nearest neighbor search for gpus. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 4236–4247. IEEE, 2024.

[55] OpenAI. Gpt-image-1. https://openai.com/zh-Hans-CN/index/introducing-4o-image-generation/, 2025.

[56] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.

[57] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[58] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025.

[59] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[60] rapidsai. cuGraph - RAPIDS Graph Analytics Library. https://github.com/rapidsai/cugraph, 2018. Accessed: 2025-11-12.

[61] Recraft. Recraft v3. https://www.recraft.ai/docs/recraft-models/recraft-V3, 2024.

[62] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[64] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

[65] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024.

[66] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[67] FLUX-Krea Team. Flux.1 krea [dev]. https://github.com/krea-ai/flux-krea, 2025.

[68] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[69] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdul-mohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[70] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. 2023.

[71] Huy V Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, et al. Automatic data curation for self-supervised learning: A clustering-based approach. *arXiv preprint arXiv:2405.15613*, 2024.

[72] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[73] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[74] Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

[75] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

[76] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[77] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025.

[78] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

[79] Xianfeng Wu, Yajing Bai, Haoze Zheng, Harold Haodong Chen, Yexin Liu, Zihao Wang, Xuran Ma, Wen-Jie Shu, Xianzu Wu, Harry Yang, et al. Lightgen: Efficient image generation through knowledge distillation and direct preference optimization. *arXiv preprint arXiv:2503.08619*, 2025.

[80] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025.

[81] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng YU, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In *Forty-second International Conference on Machine Learning*.

[82] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations*.

[83] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

[84] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

[85] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[86] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.

[87] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

[88] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024.

[89] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024.

[90] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26125–26135, 2025.

[91] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[92] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

[93] Yifu Zhang, Hao Yang, Yuqi Zhang, Yifei Hu, Fengda Zhu, Chuang Lin, Xiaofeng Mei, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Waver: Wave your way to lifelike video generation. *arXiv preprint arXiv:2508.15761*, 2025.

[94] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025.

[95] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

[96] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

[97] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogview3: Finer and faster text-to-image generation via relay diffusion. *arXiv preprint arXiv:2403.05121*, 2024.

[98] Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3dis-flux: simple and efficient multi-instance generation with dit rendering. *arXiv preprint arXiv:2501.05131*, 2025.

[99] Le Zhuo, Ruoyi Du, Xiao Han, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[100] Le Zhuo, Songhao Han, Yuandong Pu, Boxiang Qiu, Sayak Paul, Yue Liao, Yihao Liu, Jie Shao, Xi Chen, Si Liu, et al. Factuality matters: When image generation and editing meet structured visuals. *arXiv preprint arXiv:2510.05091*, 2025.

A. Prompts Used in the Report

Here we summarize the prompts/instructions used in Figure 1-3, which can be directly input into Z-Image-Turbo (with PE disabled) to reproduce our generation results.

A.1. Figure 1

Column #1

- Case #1:
一张中景手机自拍照片拍摄了一位留着长黑发的年轻东亚女子在灯光明亮的电梯内对着镜子自拍。她穿着一件带有白色花朵图案的黑色露肩短上衣和深色牛仔裤。她的头微微倾斜，嘴唇嘟起做亲吻状，非常可爱俏皮。她右手拿着一部深灰色智能手机，遮住了部分脸，后置摄像头镜头对着镜子。电梯墙壁由抛光不锈钢制成，反射着头顶的荧光灯和主体。左侧墙上有一个带有许多圆形按钮和小型数字显示屏的垂直面板。在按钮下方，可以看到一个金属扶手。后墙上贴着带有文字的长方形标志。地面铺着带有白色纹理的深色大理石瓷砖。整体照明为人造光，明亮，具有电梯内部的特征。
(Translation: A mid-range phone selfie captured a young East Asian woman with long black hair taking a selfie in front of a mirror in a brightly lit elevator. She was wearing a black off shoulder short top with a white floral pattern and dark jeans. Her head tilted slightly, and her lips curled up in a kiss, very cute and playful. She held a dark gray smartphone in her right hand, covering part of her face, with the rear camera facing the mirror. The elevator walls are made of polished stainless steel, reflecting the fluorescent lights and main body above the head. There is a vertical panel on the left wall with many circular buttons and a small digital display screen. Below the button, you can see a metal armrest. There is a rectangular sign with text on the back wall. The ground is covered with dark marble tiles with white texture. The overall lighting is artificial light, bright, and has the characteristics of an elevator interior.)

- Case #2:
一张充满动感的运动摄影照片，捕捉到一名橄榄球运动员在比赛中奔跑的瞬间，他正向右冲刺，左手抱着橄榄球。中心人物是一名30多岁的黑人男性橄榄球运动员，身穿绿白条纹球衣，上面有黄色闪电标志，胸口白色小字清晰的写着"MAI"，没有其他明显logo，白色袜子带有绿色条纹，脚穿白色钉鞋。他的左侧，两名身穿黑色和黄色条纹球衣的对手正朝相反方向奔跑，球衣正面有"Flux"字样，搭配黄色袜子和白色钉鞋，均略微失焦。背景是模糊的体育场观众席，观众穿着各色服装，还有蓝色和白色的体育场座椅，以及一块橙色广告牌，上面有部分可见的白色文字"Tongyi"和一个黄色圆形标志。前景是维护良好的绿色橄榄球场。动作摄影，体育摄影，浅景深，中心球员对焦清晰，背景虚化，自然光照明，色彩鲜艳，高对比度，构图动感，对手球员有运动模糊，充满活力，竞技氛围，户外体育场场景。
(Translation: A dynamic sports photography shot capturing a rugby player running during a game, sprinting to the right with a rugby ball tucked under his left arm. The central figure is a Black male rugby player in his 30s, wearing a green and white striped jersey with a yellow lightning bolt logo, the white text "MAI" clearly visible on the chest with no other prominent logos, white socks with green stripes, and white cleats. To his left, two opponents in black and yellow striped jerseys are running in the opposite direction, their jerseys displaying "Flux" on the front, paired with yellow socks and white cleats, both slightly out of focus. The background shows a blurred stadium crowd wearing various colored clothing, blue and white stadium seats, and an orange advertising board with partially visible white text "Tongyi" and a yellow circular logo. The foreground features a well-maintained green rugby field. Action photography, sports photography, shallow depth of field, central player in sharp focus, blurred background, natural

lighting, vibrant colors, high contrast, dynamic composition, motion blur on opponent players, energetic, competitive atmosphere, outdoor stadium scene.)

- Case #3:

一张广角、平视角度的照片捕捉到了一个充满活力的街景，地点是一条铺着不平整鹅卵石的狭窄古老小巷。小巷两旁是两到三层的红砖建筑，具有传统建筑特色的深色木门、窗框和悬挑的楼层。在左边，两名女子站在一片阳光下交谈。一名女子在格子衬衫和深色裤子外套着红色围裙，另一名女子则披着深色披肩。一只黑色小狗躺在她们脚下温暖的石头上。在前景中，一只体型较大、有着蓬松卷曲尾巴的金色狗正在嗅探鹅卵石路面。沿着小巷的中心再往前，一个人正骑着一辆小型摩托车远去，另一只黑色小狗则坐在街道的右侧。明亮的阳光和深邃的阴影在整个场景中形成鲜明对比，突显了砖块和石头的纹理。小巷尽头可见的天空是苍白的阴白色。

(Translation: A wide-angle, head up photo captures a vibrant street scene in a narrow ancient alley paved with uneven pebbles. On both sides of the alley are two to three story red brick buildings with traditional architectural features such as dark wooden doors, window frames, and cantilevered floors. On the left, two women are standing in a sunny area talking. A woman is wearing a red apron over a checkered shirt and dark pants, while another woman is draped in a dark shawl. A small black dog lay on the warm stone beneath their feet. In the foreground, a large golden dog with a fluffy and curly tail is sniffing the cobblestone road surface. Continuing along the center of the alley, a person is riding a small motorcycle away, while another black puppy is sitting on the right side of the street. The bright sunlight and deep shadows create a sharp contrast throughout the scene, highlighting the texture of the bricks and stones. The sky visible at the end of the alley is pale and gloomy white.)

- Case #4:

一张宁静的、全景横向照片，捕捉了一个小孩侧身站在郁郁葱葱的绿色草岸上，旁边是平静的水体。场景设置在黄金时刻，很可能是日落时分，色调柔和而温暖。孩子位于画面左侧，戴着一顶浅色的编织草帽，在浅蓝白格纹长袖衬衫外穿着一件橄榄绿色的短袖背心，下身是宽松的深蓝色牛仔裤，裤脚卷起，露出棕色的鞋子。孩子的右手拿着一朵黄色小花的茎，左手提着一个银色的小号镀锌金属喷壶。他/她正朝右边望去，看向水面。前景是点缀着黄色小野花的草坡。中景是河流或池塘的静水，倒映着天空温暖的粉橙色调。对岸有绿色植被和一堆灰色岩石。背景是柔和模糊的，展现了广阔的绿色田野、远处的树林线以及一些模糊的建筑轮廓，这一切都在一片广阔的天空下，天空布满了柔和的、带有粉色和橙色渐变的云彩。摄影风格的特点是浅景深，在背景中创造了显著的散景效果，使主体突出。光线自然而漫射，营造出平和、田园诗般和怀旧的氛围。孩子的脚被画面的底部边缘轻微截断。

(Translation: A peaceful, panoramic horizontal photo captures a child standing sideways on a lush green grassy bank, with a calm body of water beside it. The scene is set in prime time, most likely at sunset, with soft and warm tones. The child is located on the left side of the screen, wearing a light colored woven straw hat, an olive green short sleeved vest over a light blue and white checkered long sleeved shirt, and loose dark blue jeans with rolled up hemlines revealing brown shoes. The child holds a stem of a small yellow flower in their right hand and a silver small galvanized metal spray can in their left hand. He/she is looking to the right, towards the water surface. The prospect is a grassy slope adorned with small yellow wildflowers. The central view is the still water of a river or pond, reflecting the warm pink orange tones of the sky. There is green vegetation and a pile of gray rocks on the opposite bank. The background is soft and blurry, showing vast green fields, distant forest lines, and some blurry building contours, all under a vast sky filled with soft, pink and orange gradient clouds. The characteristic of photography style is shallow depth of field, which creates a significant bokeh effect in the background, making the subject stand out. The natural and diffuse light creates a peaceful, pastoral, and nostalgic atmosphere. The child's feet were slightly cut off by the bottom edge of the screen.)

- Case #5:

一张广角风景照片，拍摄于阴天的安徽宏村古村落。画面被平静水体的岸线水平分割，形成了村庄与天空近乎完美的镜面倒影。在中景部分，一簇密集的传统徽派建筑沿水边排列，具有独特的白墙和深灰色瓦顶。几栋建筑的屋檐下悬挂着红色的纸灯笼，在柔和的背景中增添了鲜艳的色彩点缀。水边的石板路上和房屋之间散布着许多光秃秃的落叶树。一些身影微小的人沿着这条小路行走或坐着。在背景中，一片朦胧的蓝绿色山脉在淡灰色的天空下连绵起伏。右侧山坡上可以看到一个小型输电塔。在画面中心偏右的一栋建筑上，门楣上方挂着一块横向的木匾，上面有黑色的汉字"世德堂"。该摄影作品的风格特点是构图对称，光线柔和漫射，景深较大，整个场景都清晰锐利，色调清冷而宁静，以白色、灰色和蓝色为主，红色作为强烈的点缀色。整体氛围平和、安详且具有永恒

感。

(Translation: A wide-angle landscape photo taken on a cloudy day in the ancient village of Hongcun, Anhui. The screen is horizontally divided by the calm shoreline of the water, forming a nearly perfect mirror reflection of the village and the sky. In the central area, a dense cluster of traditional Huizhou style buildings are arranged along the water's edge, featuring unique white walls and dark gray tiled roofs. Red paper lanterns hang under the eaves of several buildings, adding vibrant color accents to the soft background. There are many bare deciduous trees scattered between the stone roads and houses by the water's edge. Some small figures walked or sat along this path. In the background, a hazy blue-green mountain range undulates continuously under a light gray sky. A small transmission tower can be seen on the right slope. On a building to the right of the center of the screen, there is a horizontal wooden plaque hanging above the lintel, with black Chinese characters "世德堂" on it. The stylistic features of this photography work are symmetrical composition, soft and diffuse lighting, large depth of field, clear and sharp entire scene, cool and peaceful tones, mainly white, gray and blue, with red as a strong accent color. The overall atmosphere is peaceful, serene, and has a sense of eternity.)

- Case #6:
一张充满活力的广角夜景照片，捕捉了中国广州猎德大桥上空壮观的烟花表演。场景设置在漆黑的夜空下，被多朵巨大的烟花爆炸瞬间照亮。烟花主要为白色和红色/粉色，在画面的上半部分形成了灿烂的牡丹花状图案，周围环绕着浓浓的硝烟。猎德大桥，一座现代化的斜拉桥，在中景处横跨珠江。其独特的拱形中央桥塔被温暖的黄色灯光照亮。在这个中央桥塔的正面，可以看到一个被部分遮挡的红色小标志。桥面也被路灯照亮。在前景中，黑暗的江水映照出烟花和桥灯的缤纷倒影。左下角可以看到一艘小船的黑色剪影，远处还散布着其他更小的船只。背景是闪闪发光的现代化城市天际线，摩天大楼和其他建筑上的无数灯光点缀其间。该摄影风格以长曝光为特点，这从烟花的轨迹中可以明显看出，营造出一种动感和喜庆的氛围。图像对比度高，对焦清晰，在黑暗的环境中呈现出鲜艳的色彩。

(Translation: A vibrant wide-angle night view photo captures the spectacular fireworks display over the Liede Bridge in Guangzhou, China. The scene is set in the pitch black night sky, instantly illuminated by multiple huge fireworks explosions. The fireworks are mainly white and red/pink, forming a brilliant peony shaped pattern in the upper part of the picture, surrounded by thick gunpowder smoke. Liede Bridge, a modern cable-stayed bridge, crosses the the Pearl River in the middle view. Its unique arched central bridge tower is illuminated by warm yellow lights. On the front of this central bridge tower, a partially obscured red small sign can be seen. The bridge deck is also illuminated by streetlights. In the foreground, the dark river reflects the colorful reflections of fireworks and bridge lights. In the lower left corner, a black silhouette of a small boat can be seen, with other smaller boats scattered in the distance. The background is a sparkling modern city skyline, adorned with countless lights from skyscrapers and other buildings. This photography style is characterized by long exposures, which can be clearly seen from the trajectory of fireworks, creating a dynamic and festive atmosphere. The image has high contrast, clear focus, and presents bright colors in dark environments.)

Column #2

- Case #1:
A stylish young woman sits casually on an unmade bed bathed in soft daylight, wearing a pastel yellow oversized T-shirt with subtle white text and cozy light gray sweatpants. Her skin glows fresh beneath glossy deep lavender hydrogel under-eye patches, while her hair is tied back loosely with a scrunchie, complemented by delicate gold hoop earrings. Nearby, a tube of hand cream and an open laptop rest casually atop soft, slightly rumpled sheets. The natural window light gently illuminates her radiant skin and the subtle sheen of the hydrogel patches, enhancing the cozy textures of her loungewear and bedding. Shot from a top-down selfie angle, the framing captures her face, shoulders, and upper torso with realistic iPhone grain, conveying an authentic, relaxed self-care morning moment in a softly lit bedroom scene – skincare selfie, shot on iPhone.
- Case #2:
一张逼真的年轻东亚女性肖像，位于画面中心偏左的位置，带着浅浅的微笑直视观者。她身着以浓郁的红色和金色为主的传统中式服装。她的头发被精心盘起，饰有精致的红色和金色花卉和叶形发饰。她的眉心之间额头上绘有一个小巧、华丽的红色花卉图案。她左手持一把仿古扇子，扇面上绘

55

有一位身着传统服饰的女性、一棵树和一只鸟的场景。她的右手向前伸出，手掌向上，托着一个悬浮的发光的霓虹黄色闪电亚克力灯牌，这是画面中最亮的元素。背景是模糊的夜景，带有暖色调的人工灯光，一场户外文化活动或庆典。在远处的背景中，她头部的左侧略偏，是一座高大、多层、被暖光照亮的西安大雁塔。中景可见其他模糊的建筑和灯光，暗示着一个繁华的城市或文化背景。光线是低调的，闪电符号为她的脸部和手部提供了显著的照明。整体氛围神秘而迷人。人物的头部、手部和上半身完全可见，下半身被画面底部边缘截断。图像具有中等景深，主体清晰聚焦，背景柔和模糊。色彩方案温暖，以红色、金色和闪电的亮黄色为主。

(Translation: A realistic portrait of a young East Asian woman, located to the left of the center of the image, looking directly at the viewer with a faint smile. She was dressed in traditional Chinese clothing dominated by rich red and gold colors. Her hair was carefully styled, adorned with delicate red and gold flowers and leaf shaped hair accessories. There is a small and gorgeous red floral pattern painted on her forehead between her eyebrows. She held an antique style fan in her left hand, with a scene of a woman dressed in traditional clothing, a tree, and a bird painted on the fan surface. Her right hand extended forward, palm up, holding a suspended glowing neon yellow lightning acrylic light tag, which was the brightest element in the picture. The background is a blurry night scene with warm toned artificial lighting, representing an outdoor cultural event or celebration. In the distant background, to the left of her head is a tall, multi-layered, warm lit Xi'an Big Wild Goose Pagoda. Other blurry buildings and lights can be seen in the middle of the scene, implying a bustling city or cultural background. The light is low-key, and the lightning symbol provides significant illumination for her face and hands. The overall atmosphere is mysterious and charming. The head, hands, and upper body of the character are fully visible, while the lower body is cut off by the bottom edge of the screen. The image has a moderate depth of field, the subject is clearly focused, and the background is soft and blurry. The color scheme is warm, with red, gold, and bright yellow of lightning as the main colors.)

- Case #3:
A full-body, eye-level photograph of a young, beautiful East Asian woman posing cheerfully inside a brightly lit LEGO store or brand exhibition space. The woman, positioned slightly right of center, has long dark hair and is smiling at the camera. She wears a vibrant yellow ribbed beanie, a white diamond-quilted puffer jacket over a white t-shirt, and medium-wash blue jeans with cuffs rolled up at the ankles. She is wearing white lace-up sneakers and white socks, with a small red heart visible on her left sock. In her left hand, she holds a black structured handbag. Her pose is playful, with her left leg kicked up behind her. To her left is a large, multi-tiered display stand in bright yellow, which features the official LEGO logo – white text in a red square with a black and yellow outline – in the upper left corner. On this stand are two large-scale LEGO Minifigure statues: a policeman in a blue uniform and hat stands in the foreground, and behind him is a Santa Claus figure in red. The background shows more yellow retail shelving stocked with various LEGO sets and products. The floor is made of large, light grey tiles, and a white dome security camera is visible on the ceiling. The image is a sharp, well-lit snapshot with a vibrant color palette, dominated by yellow, red, and blue, creating a joyful and commercial atmosphere.

- Case #4:
A candid mid-2010s-style snapshot featuring a pale young woman with icy platinum hair styled casually loose, seated on a metal bench inside a monochrome concept store. She wears a huge black hoodie, sheer tights, and maroon platform creepers, complemented by a beanie embroidered with "Z-Image Real & Fast" The subject's relaxed expression gazes off to the side, conveying subtle, ambiguous emotion. The lighting is cold and matte with soft shadows stretching along a wooden floor, intentionally exhibiting muted color saturation, softened contrast, and distinctly cool-toned bluish-gray shadows. Visible textures include realistic skin details, detailed fabric grain of the hoodie and tights, individual icy hair strands, and clear accessory textures. The framing is slightly off-center and casually tilted, capturing spontaneous intimacy and informal snapshot aesthetics characteristic of mid-2010s casual youth photography.

Column #3

- Case #1:
一位男士和他的贵宾犬穿着配套的服装参加狗狗秀，室内灯光，背景中有观众。

(Translation: A man and his poodle participated in a dog show wearing matching costumes, with

indoor lighting and an audience in the background.)

- Case #2:

  一张特写、逼真的东亚婴儿肖像，婴儿穿着一件印有心形图案的奶油色蓬松冬季连体衣，直视观者。婴儿拥有深色头发和红扑扑的脸颊。婴儿手边部分可见一个色彩鲜艳的玩具，背景模糊处有一位穿着格子衬衫的人。室内光线具有柔和的阴影和高光，营造出温暖的色调，婴儿脸部清晰聚焦，背景柔和模糊。低饱和度、颗粒感、老胶片风格。

  (Translation: A close-up, realistic portrait of an East Asian baby wearing a creamy fluffy winter jumpsuit with a heart-shaped pattern, looking straight at the viewer. Babies have dark hair and rosy cheeks. A brightly colored toy can be seen near the baby's side, with a person wearing a checkered shirt in a blurry background. The indoor lighting features soft shadows and highlights, creating a warm tone. The baby's face is clearly focused, and the background is soft and blurry. Low saturation, graininess, and vintage film style.)

- Case #3:

  北京国家体育场（鸟巢）的照片，蓝天背景下，体育场的外观由复杂的交织钢结构形成网状图案主导。前景中一个人穿着休闲装，略微偏中心位置行走。背景通过钢结构可以看到体育场内部的红色座位区。"A30"用红色标记在钢结构的左下角。图像从低角度拍摄，突显建筑的宏伟和规模。照片，高对比度，戏剧性光线，蓝天，低角度视角，建筑摄影，聚焦清晰，现代设计，精细钢结构，鲜艳红色点缀，视觉冲击力强，构图平衡。

  (Translation: A photo of the Beijing National Stadium (Bird's Nest), with a blue sky background, the appearance of the stadium is dominated by a complex interwoven steel structure forming a mesh pattern. In the foreground, a person is wearing casual clothing and walking slightly off center. The background shows the red seating area inside the stadium through the steel structure. "A30" is marked in red on the bottom left corner of the steel structure. The image is taken from a low angle to highlight the grandeur and scale of the building. Photos, high contrast, dramatic lighting, blue sky, low angle perspective, architectural photography, clear focus, modern design, fine steel structure, bright red accents, strong visual impact, balanced composition.)

A.2. Figure 2

Row #1

- Case #1:

  杂志封面设计。 文案：大标题"「造相」Z-Image"。 小标题："Winter Release. Spring for Generative Art."。 版本号：" VOL 1.0"。 中间底部极小字"通义多模态交互出版社"。 拉开一片白雪茫茫下的拉链，拉链下漏出绿草鲜花的春天，移轴微距，拉链是一个冒着白烟远去的火车头，精美构图，夸张的俯视视角，视觉冲击力，高对比度，高饱和度。

  (Translation: Magazine cover design. Copy: Headline "「造相」Z-Image". Subtitle: "Winter Release. Spring for Generative Art. Version number: "VOL 1.0". The extremely small font at the bottom of the middle reads '通义多模态交互出版社'. Pulling open a zipper under a vast expanse of white snow, the spring of green grass and flowers peeks out from under the zipper. Moving the axis macro, the zipper is a locomotive emitting white smoke far away, with exquisite composition, exaggerated top-down perspective, visual impact, high contrast, and high saturation.)

- Case #2:

  一幅垂直构图、风格化的数字插画，设计为一张励志海报。场景描绘了夜间的沙漠景观，头顶是广阔无垠、繁星密布的天空，其中银河清晰可见。前景和中景以深蓝色近乎黑色的剪影为特色。左侧，一棵巨大而细节丰富的约书亚树剪影占据了画面主导。更远处可以看到两棵较小的约书亚树。右侧，两个人的剪影站在一个小山丘上，仰望着天空。天空从底部的深海军蓝过渡到顶部的浅蓝色，布满星辰，明亮的银河带以柔和的白色、紫色和蓝色调，从右上角划过。图像上覆盖有五处独立的渲染中文字样。顶部是白色大号字体，内容为"于无垠黑暗中，寻见你的微光"。在中间靠近人物的位置，有较小的黑色字体写着"心之所向，宇宙回响"。在最底部，是白色大号艺术字的主标题"仰望·逐梦"，其下方是稍小的白色字体"心的旅程由此开始"。在山丘上靠近人物的地方，有一个非常小、几乎隐藏的黑色签名"观星者"。整体风格图形化且简约，将扁平的剪影与细节更丰富、富有绘画感的天空相结合，营造出一种深沉、引人深思且充满希望的氛围。

  (Translation: A vertically composed and stylized digital illustration designed as an inspirational poster. The scene depicts a desert landscape at night, with a vast and starry sky overhead, among which the Milky Way is clearly visible. The foreground and middle ground are characterized by

deep blue and almost black silhouettes. On the left, a large and detailed silhouette of a Joshua tree dominates the scene. Two smaller Joshua trees can be seen further away. On the right, silhouettes of two people stand on a small hill, looking up at the sky. The sky transitions from deep sea navy blue at the bottom to light blue at the top, filled with stars, and the bright Milky Way streaks across in soft white, purple, and blue tones from the top right corner. The image is covered with five independent rendered Chinese characters. At the top is a large white font that reads '于无垠黑暗中，寻见你的微光'. In the middle, near the character, there is a small black font that reads '心之所向，宇宙回响'. At the bottom, there is the main title "仰望·逐梦" in large white artistic font, and below it is a slightly smaller white font "心的旅程由此开始". Near the character on the hill, there is a very small, almost hidden black signature called '观星者'. The overall style is graphical and minimalist, combining flat silhouettes with a more detailed and picturesque sky, creating a deep, thought-provoking, and hopeful atmosphere.)

- Case #3:

  一张充满活力的视觉作品集平面设计海报，整张图片以非常小的透明棋盘格为背景，展示了一个3D渲染的卡通人物。画面左侧是一位年轻女性的半身像，她皮肤白皙，留着深棕色长卷发，戴着粉色边框的眼镜，眼镜后是棕色的大眼睛。她笑容灿烂，露出牙齿，戴着小巧的银色耳钉。她的着装包括一件浅灰色西装外套、一件白色翻领衬衫和一条红色领带。她手中捧着一束由四朵鲜艳的黄色向日葵组成的花束，花茎为绿色。该角色被一圈粗白的轮廓线包围，使其从背景中凸显出来。海报的右侧主要是大型艺术字。主标题"视觉作品集"采用粗大的黄色笔刷风格字体。其上叠加着一行纤细的红色草书英文"Personalization"。下方是圆润气泡状的黄色小一号字体"VISUAL PORTFOLIO"。其下写出了三个亮点："·中英渲染，字字如刻 ● Bilingual Rendering" "·不止真实，更懂美学 ● Realism & Aesthetic" "·读懂复杂，生成精妙 ● Complexity & Elegance" 这里中文是白色手写体大字，英文是半透明的印刷体小字。海报包含多个文本块和标志。中上部先是黄色的文字"Z-Image x"，中间是一个戴着耳机的卡通头像的黄色线条画标志，后面跟着文字"x Design"。在右下角有一个可爱的拟人化扩音器，它有两只大大的眼睛，颜色为浅绿色和奶油色，底部有一朵小雏菊。整体风格是3D角色渲染和平面设计的结合，特点是氛围愉快、对比度高，并采用了黄、黑、白为主的配色方案。

  (Translation: A vibrant visual portfolio graphic design poster, with a very small transparent checkerboard pattern as the background, showcasing a 3D rendered cartoon character. On the left side of the screen is a half body portrait of a young woman with fair skin, long curly dark brown hair, wearing pink framed glasses, and brown big eyes behind the glasses. She had a bright smile, revealing her teeth and wearing small silver earrings. Her attire includes a light gray suit jacket, a white collared shirt, and a red tie. She held a bouquet of four bright yellow sunflowers in her hand, with green stems. The character is surrounded by a thick white outline, making it stand out from the background. On the right side of the poster are mainly large artistic characters. The main title "Visual Works Collection" adopts a thick yellow brush style font. On top of it is a thin line of red cursive English word "Personalization". Below is a round, bubble shaped yellow font with one size smaller reading 'VISUAL PORTFOLIO'. Below are three highlights: "·中英渲染，字字如刻 ● Bilingual Rendering" "·不止真实，更懂美学 ● Realism & Aesthetic" "·读懂复杂，生成精妙 ● Complexity & Elegance" The Chinese characters here are white handwritten large characters, while the English characters are semi transparent printed small characters. The poster contains multiple text blocks and logos. The upper part is first marked with yellow text "Z-Image x", followed by a yellow line drawn logo of a cartoon avatar wearing headphones, and then the text "x Design". In the bottom right corner, there is a cute anthropomorphic amplifier with two big eyes in light green and cream colors, and a small daisy at the bottom. The overall style is a combination of 3D character rendering and graphic design, characterized by a pleasant atmosphere, high contrast, and predominantly yellow, black, and white color schemes.)

Row #2

- Case #1:

  一张虚构的英语电影《回忆之味》（The Taste of Memory）的电影海报。场景设置在一个质朴的19世纪风格厨房里。画面中央，一位红棕色头发、留着小胡子的中年男子（演员阿瑟·彭哈利根饰）站在一张木桌后，他身穿白色衬衫、黑色马甲和米色围裙，正看着一位女士，手中拿着一大块生红肉，下方是一个木制切菜板。在他的右边，一位梳着高髻的黑发女子（演员埃莉诺·万斯饰）倚靠在桌子上，温柔地对他微笑。她穿着浅色衬衫和一条上白下蓝的长裙。桌上除

了放有切碎的葱和卷心菜丝的切菜板外，还有一个白色陶瓷盘、新鲜香草，左侧一个木箱上放着一串深色葡萄。背景是一面粗糙的灰白色抹灰墙，墙上挂着一幅风景画。最右边的一个台面上放着一盏复古油灯。海报上有大量的文字信息。左上角是白色的无衬线字体"ARTISAN FILMS PRESENTS"，其下方是"ELEANOR VANCE"和"ACADEMY AWARD® WINNER"。右上角写着"ARTHUR PENHALIGON"和"GOLDEN GLOBE® AWARD WINNER"。顶部中央是圣丹斯电影节的桂冠标志，下方写着"SUNDANCE FILM FESTIVAL GRAND JURY PRIZE 2024"。主标题"THE TASTE OF MEMORY"以白色的大号衬线字体醒目地显示在下半部分。标题下方注明了"A FILM BY Tongyi Interaction Lab"。底部区域用白色小字列出了完整的演职员名单，包括"SCREENPLAY BY ANNA REID"、"CULINARY DIRECTION BY JAMES CARTER"以及Artisan Films、Riverstone Pictures和Heritage Media等众多出品公司标志。整体风格是写实主义，采用温暖柔和的灯光方案，营造出一种亲密的氛围。色调以棕色、米色和柔和的绿色等大地色系为主。两位演员的身体都在腰部被截断。

(Translation: A movie poster for the fictional English movie 'The Taste of Memory'. The scene is set in a rustic 19th century style kitchen. In the center of the screen, a middle-aged man with reddish brown hair and a small beard (played by actor Arthur Penhaligan) stands behind a wooden table. He is wearing a white shirt, black vest, and beige apron, looking at a woman holding a large piece of raw red meat with a wooden cutting board below. On his right, a black haired woman with a high bun (played by actress Eleanor Vance) leaned against the table and smiled gently at him. She was wearing a light colored shirt and a long skirt with white on top and blue on the bottom. On the table, in addition to a chopping board with chopped onions and shredded cabbage, there is also a white ceramic plate and fresh herbs. On the left side, there is a wooden box with a string of dark grapes. The background is a rough gray white plaster wall with a landscape painting hanging on it. On the far right countertop is a vintage oil lamp. There is a lot of textual information on the poster. The white sans serif font "ARTISAN FILMS PRESS" is located in the upper left corner, with "ELEANOR VANCE" and "ACADEMY AWARD" below it ® WINNER". In the upper right corner are written "ARTHUR PENHALIGON" and "GOLDEN GLOBE" ® AWARD WINNER". At the top center is the crown emblem of Sundance Film Festival, with the words' SUNDANCE FILM FESTIVAL GRAND JURY PRIZE 2024 'written below. The main title "THE TASTE OF Memory" is prominently displayed in large white serif font in the lower half. The title reads 'A FILM BY Tongyi Interaction Lab.'. The bottom area lists the complete cast and crew list in small white font, including "SCREENPLAY BY ANNA REID", "CULINARY Directing BY JAMES CARTER", as well as many production company logos such as Artisan Films, Riverstone Pictures, and Heritage Media. The overall style is realism, using warm and soft lighting schemes to create an intimate atmosphere. The color scheme is dominated by earthy tones such as brown, beige, and soft green. The bodies of both actors were severed at the waist.)

- Case #2:
  一张竖版日本艺术展海报，背景为深蓝色。设计以醒目的黄色文字和七幅水彩画拼贴为主。顶部是日文和英文标题。日文部分使用大号黄色宋体风格字体，内容为"谷川 正季 - 水彩画の世界 -"。其下方是较小的黄色无衬线字体"-The world of watercolor-"。主标题"中天堂"以非常大的风格化黄色字体突出显示。其下是英文翻译"HEAVEN OF DREAM"，同样为黄色无衬线字体。再下一行是日文副标题"我が心の桂林"，字体较大，后跟其英文翻译"GUILIN IN MY MIND"，字体较小。海报中央是由七幅描绘桂林喀斯特地貌不同场景的水彩画组成的网格。这些画作展示了云雾缭绕的群山、蜿蜒穿过山谷的河流、倒映在水面上的绚丽日落、人们在船上提着灯笼的夜景以及其他富有氛围的风景。海报底部三分之一处用较小的黄色文字列出了活动详情，包括"2025.11.11(六) 17(五) 9:00 20:00"，"阿里巴巴云谷园区"，"(021)-34567890"。整体风格是优雅的平面设计，采用了深蓝色和黄色的高对比度双色调色板。

  (Translation: A vertical Japanese art exhibition poster with a dark blue background. The design mainly features eye-catching yellow text and seven watercolor collages. At the top are Japanese and English titles. The Japanese section uses large yellow Song style fonts and the content is "谷川 正季 - 水彩画の世界 -". Below it is a smaller yellow sans serif font that reads' The world of watercolor - '. The main title "中天堂" is highlighted in a very large stylized yellow font. Below is the English translation "HEAVEN OF DREAM", also in yellow sans serif font. The next line is the Japanese subtitle '我が心の桂林', with a larger font size, followed by its English translation 'GUILIN IN MY MIND', with a smaller font size. In the center of the poster is a grid composed of seven watercolor paintings depicting different scenes of Guilin's karst landscape. These paintings showcase misty mountains, winding rivers through valleys, stunning sunsets reflected on the water, night scenes of

people carrying lanterns on boats, and other atmospheric landscapes. The activity details are listed in small yellow text at the bottom third of the poster, including "2025.11.11(六) 17(五) 9:00 20:00", "阿里巴巴云谷园区", and "(021) -34567890". The overall style is an elegant graphic design, featuring a high contrast dual tone palette of dark blue and yellow.)

- Case #3:
一张方形构图的特写照片，主体是一片巨大的、鲜绿色的植物叶片，并叠加了文字，使其具有海报或杂志封面的外观。主要拍摄对象是一片厚实、有蜡质感的叶子，从左下角到右上角呈对角线弯曲穿过画面。其表面反光性很强，捕捉到一个明亮的直射光源，形成了一道突出的高光，亮面下显露出平行的精细叶脉。背景由其他深绿色的叶子组成，这些叶子轻微失焦，营造出浅景深效果，突出了前景的主叶片。整体风格是写实摄影，明亮的叶片与黑暗的阴影背景之间形成高对比度。图像上有多处渲染文字。左上角是白色的衬线字体文字"PIXEL-PEEPERS GUILD Presents"。右上角同样是白色衬线字体的文字"[Instant Noodle] 泡面调料包"。左侧垂直排列着标题"Render Distance: Max"，为白色衬线字体。左下角是五个硕大的白色宋体汉字"显卡在...燃烧"。右下角是较小的白色衬线字体文字"Leica Glow™ Unobtanium X-1"，其正上方是用白色宋体字书写的名字"蔡几"。
(Translation: A close-up photo with a square composition, featuring a large, bright green plant leaf and overlaid with text to give it the appearance of a poster or magazine cover. The main subject being photographed is a thick, waxy leaf that curves diagonally through the frame from the bottom left corner to the top right corner. Its surface has strong reflectivity, capturing a bright direct light source and forming a prominent highlight, revealing parallel fine leaf veins under the bright surface. The background is composed of other dark green leaves that are slightly out of focus, creating a shallow depth of field effect and highlighting the main leaf of the foreground. The overall style is realistic photography, with high contrast between bright leaves and dark shadow backgrounds. There are multiple rendered texts on the image. In the upper left corner is the white serif font text "PIXEL-PEEPERS GUIDE Gifts". The text in white serif font in the upper right corner reads '[Instant Noodle] 泡面调料包'. The title "Render Distance: Max" is vertically arranged on the left side in white serif font. In the bottom left corner are five large white Song typeface Chinese characters that read '显卡在...燃烧'. The smaller white serif font text "Leica Glow" is located in the bottom right corner ™ Unobtanium X-1"，Above it is the name "蔡几" written in white Song typeface.)

Row #3

- Case #1:
A vertical digital illustration depicting a serene and majestic Chinese landscape, rendered in a style reminiscent of traditional Shanshui painting but with a modern, clean aesthetic. The scene is dominated by towering, steep cliffs in various shades of blue and teal, which frame a central valley. In the distance, layers of mountains fade into a light blue and white mist, creating a strong sense of atmospheric perspective and depth. A calm, turquoise river flows through the center of the composition, with a small, traditional Chinese boat, possibly a sampan, navigating its waters. The boat has a bright yellow canopy and a red hull, and it leaves a gentle wake behind it. It carries several indistinct figures of people. Sparse vegetation, including green trees and some bare-branched trees, clings to the rocky ledges and peaks. The overall lighting is soft and diffused, casting a tranquil glow over the entire scene. Centered in the image is overlaid text. At the top of the text block is a small, red, circular seal-like logo containing stylized characters. Below it, in a smaller, black, sans-serif font, are the words 'Zao-Xiang * East Beauty & West Fashion * Z-Image'. Directly beneath this, in a larger, elegant black serif font, is the word 'SHOW & SHARE CREATIVITY WITH THE WORLD'. Among them, there are "SHOW & SHARE", "CREATIVITY", and "WITH THE WORLD"

- Case #2:
vertical movie poster for the film "Come Back Home Often." created by Master of Oil painting. The artwork is a unified digital painting with a heavy impasto texture, mimicking thick oil paint strokes applied with a palette knife. The central focus is a massive, abstract figure rendered in thick, textured white paint, resembling a giant bird or a stylized human form. This white shape is set against a dark navy blue background that is densely covered with small, stylized flowers painted in vibrant red and white, with green stems. In the bottom right corner, two elderly people are depicted from behind, walking away from the viewer. One person, slightly ahead, wears a purple jacket and uses a wooden cane. The other, slightly behind, wears a greyish-blue jacket. Their

bodies are truncated at the ankles by the bottom edge of the frame. The overall style is surreal and symbolic, with a high-contrast color palette dominated by deep navy, white, and red. Text control: all lettering is fully integrated into the painted surface with identical heavy impasto, each character exhibiting raised ridges and knife-scraped edges that catch ambient light. In the top left corner, in white sans-serif strokes sculpted with thick, palette-knife ridges, the words "Z-Image" appear, and directly beneath them, still in raised impasto, "Visionary Creator." In the bottom left, the Chinese title is rendered in large, white, cursive calligraphy (草书 style) built up from layered, knife-pressed paint: "常回家看看", its down-strokes showing visible paint peaks. Below this, in a smaller white serif font whose letterforms are similarly embossed with raised impasto, reads the English title: "Come Back Home Often."

- Case #3:

传统中国水墨画照片，描绘了萧瑟秋日黄昏景象，位于页面左侧。画作竖向排列，用枯笔勾勒盘曲老藤缠绕古树，浓墨点染栖息昏鸦，淡墨晕染暮色天空。枯藤如龙蛇般攀附在虬曲树干上，三两只乌鸦停驻枝头，剪影般的轮廓。远处隐约可见小桥流水和古朴人家屋舍，炊烟袅袅。近景有一条黄土古道，瘦马低头缓行。画面上方有几行黑色行书书法，题写元曲名句，右上角钤一枚朱红方印。在画作右侧，有两列竖排中文文字，纯白色背景，草书字体，第一列写着 "天净沙·秋思"，第二列写着"枯藤老树昏鸦，小桥流水人家"。传统中式文人画风格，水墨单色系带飞白枯笔效果，笔触苍劲有力，构图疏密有致，留白处理突出，充满萧瑟苍凉的美学意蕴，氛围孤寂而悠远，具有浓厚的古典诗意和文化韵味。

(Translation: The traditional Chinese ink painting photo depicts a bleak autumn dusk scene, located on the left side of the page. The painting is arranged vertically, using a dry brush to outline the winding old vines around the ancient trees, with thick ink coloring the roosting crows, and light ink blending the twilight sky. The withered vine clung to the winding tree trunk like a dragon or snake, with three or two crows perched on the branches, forming a silhouette like silhouette. In the distance, small bridges, flowing water, and quaint houses can be faintly seen, with smoke rising from cooking. In the close-up, there is a loess ancient road, and thin horses are walking slowly with their heads down. There are several lines of black cursive calligraphy above the screen, inscribed with famous lines of Yuan opera, and a vermilion square seal is stamped in the upper right corner. On the right side of the painting, there are two vertical columns of Chinese characters with a pure white background and cursive font. The first column reads "天净沙·秋思", and the second column reads "枯藤老树昏鸦，小桥流水人家". The traditional Chinese literati painting style features a single color ink wash with a flying white and withered pen effect. The brushstrokes are vigorous and powerful, the composition is dense and orderly, and the white space treatment is prominent. It is full of desolate and desolate aesthetic connotations, creating a lonely and distant atmosphere with strong classical poetry and cultural charm.)

Row #4

- Case #1:

竖排行书书法作品特写，以米白色带浅淡肌理的半生熟宣纸为载体，纸张带有自然轻微褶皱，质感温润柔和；黑色墨汁书写的行书字体，笔锋粗细变化灵动，如"人"字捺笔舒展劲挺、"志"字连笔婉转流畅，墨色层次丰富，部分笔画带自然飞白效果，尽显苍劲洒脱的书写张力；文字按传统从右至左竖列排布，可见"人扶我青志""我自踏雪至山巅"等词句，左侧配有"造相大师"落款小字，纸面点缀多枚朱红方形篆刻印章，印泥色泽饱满、印文线条清晰；多张书法纸呈轻微重叠的错落摆放，背景隐约露出其他纸张的淡色字迹，营造出随性的创作氛围；光线为柔和自然光，均匀铺洒在纸面，凸显墨色的光泽与纸张的纹理褶皱，造梦师的诗意视觉风格，整体氛围雅致古朴，兼具手写书法的灵动随性与传统文房的沉静质感。

(Translation: Close up of vertical cursive calligraphy works, using semi ripe rice paper with a light texture and off white color as the carrier. The paper has natural slight wrinkles and a warm and soft texture; The running script font written in black ink has dynamic changes in stroke thickness, such as the stretching and vigorous strokes of the "人" character and the smooth and graceful strokes of the "志" character. The ink layers are rich, and some strokes have a natural flying white effect, showcasing the vigorous and free spirited writing tension; The text is arranged vertically from right to left according to tradition, with phrases such as "人扶我青志" and "我自踏雪至山巅" visible. On the left side, there is a small signature of "造相大师", and the paper is decorated with multiple vermilion square seal seals. The ink color is full and the lines of the seal

are clear; Multiple calligraphy papers are arranged in a slightly overlapping and staggered manner, with the background faintly revealing the light colored handwriting of other papers, creating a casual creative atmosphere; The light is soft natural light, evenly spread on the paper surface, highlighting the luster of ink color and the texture wrinkles of the paper. The poetic visual style of the dream maker creates an elegant and rustic atmosphere, combining the agility and casualness of handwritten calligraphy with the calm texture of traditional study rooms.)

- Case #2:
  一张垂直构图的平面设计海报，背景是纯粹而鲜艳的宝蓝色。顶部的巨大无衬线字体主标题，上半部分为浅灰色的"Sofa Montain Slummerfest"，下半部分为白色的"Annual Napping Festival 2025"。其下方是巨大的黑色书法字体中文标题"沙发山打呼节"。海报的下半部分由一幅巨大的、插画风格的老虎插画占据，它正趴着面向观众，眼睛是黄色的。其皮毛由橙、黑、白三色构成。一个包含红色爱心的思想泡泡漂浮在它的头顶。海报上布满了详尽的活动文字。左栏用白色字体列出了以猫为主题的乐队名，如"The Fluffy Paws Grumbers (毛爪咕噜)"、"DJ Meow Mix"、"九命怪猫 (Nine Lives)"、"激光笔追逐者 (The Laser Dots)"、"纸箱爱好者 (Cardbock Box Lovers)"、"呼噜神教 (The Purr-fectionists)"、"猫草成瘾者 (The Catnip Junkies)"、"DJ Chairman Meow (猫主席)"以及像"Varh Radator Fesidenl Paw-Five"这样的无意义短语。右栏列出了活动细节，其中许多都带有滑稽的拼写错误或无意义内容，包括日期"4/1 MONDAY SUNL SUNSET"、地点"上海市浦东新区猫抓板路1号顶楼阳台"以及票务信息如"ADV. 1 CAN OF TUNA, DOOR 2 CANS, KITTENS FREE!"。在最底部是一排虚构的赞助商标志，名称包括"Catberd"、"好主人罐罐有限公司 (Good Oinar Canned Food Ltd)"和"iNONEPAWS"。

  (Translation: A vertically composed graphic design poster with a pure and vibrant navy blue background. The main title features a large sans serif font at the top, with the upper half in light gray reading "Sofa Montain Slummerfest" and the lower half in white reading "Annual Napping Festival 2025". Below it is a huge black calligraphy font with the Chinese title "沙发山打呼节". The lower part of the poster is occupied by a huge, illustrated tiger illustration, which is lying face down to the audience with yellow eyes. Its fur is composed of three colors: orange, black, and white. A thought bubble containing a red heart floats above its head. The poster is filled with detailed activity text. The left column lists the names of cat themed bands in white font, such as "The Fluffy Paws Grumbers (毛爪咕噜)", "DJ Meow Mix", "九命怪猫 (Nine Lives)", "激光笔追逐者 (The Laser Dots)", "纸箱爱好者 (Cardbock Box Lovers)", "呼噜神教 (The Purr-fectionists)", "猫草成瘾者 (The Catnip Junkies)", "DJ Chairman Meow (猫主席)" and meaningless phrases like "Varh Radator Feseidel Paw Five. The right column lists the details of the event, many of which have humorous spelling errors or meaningless content, including the date "4/1 MONDAY SUNL SUNSET", location "上海市浦东新区猫抓板路1号顶楼阳台", and ticketing information such as "ADV. 1 CAN OF TUNA, DOOR 2 CANS, KITTENS FREE. At the bottom is a row of fictional sponsor logos, with names including "Cattered," "好主人罐罐有限公司" and "iNONEPAWS".)

- Case #3:
  一张植物展览的平面设计海报，背景为素净的米白色。海报上有多幅水彩插画，描绘了各种苔藓和蕨类植物，用色以绿色、棕色和黄色为主，并配有精致的黑色墨水轮廓线。画面中央是一幅巨大而精细的绿色地钱插画，上面有棕色的孢子体。其他较小的插图散布在周围，包括左上角的细叶满江红（Fissidens），顶部中央的巨叶红茎藓（Rhodobryum giganteum），右上角的青苔属（Bryum sp.），以及右下角的凤尾藓（Marchantia formosana）。海报布局干净、简约，文字垂直和水平排列。右上角是纵向排列的黑色宋体大字"苔痕"，下边是横向排列的无衬线字体英文"Moss Exhibition"。左侧是黑色无衬线字体"Elkhorn Fern LifeStyle"。左下角写着"Alishan Moss Ecological"。日期和时间在底部突出显示："2001"为黑色大号衬线字体，其后是较小的无衬线字体"04.22 [Apr.] am. 09:00"和"05.22 [May] pm. 17:00"。每幅植物插图都附有其学名，使用小号灰色无衬线字体书写，例如"Fissidens"、"Rhodobryum giganteum"、"Bryum sp."、"Bartramiaceae"、"Alishan Moss Ecological"、"Marchantia formosana"和"Astrocella yoshinagana"。整体风格优雅且具有教育意义，将科学插画与现代排版相结合。

  (Translation: A graphic design poster for a plant exhibition, with a plain beige background. The poster features multiple watercolor illustrations depicting various mosses and ferns, primarily using green, brown, and yellow colors, with delicate black ink outlines. In the center is a large and intricate illustration of green liverwort with brown sporophytes on top. Other smaller illustrations are scattered around, including Fissidens in the upper left corner, Rhodobryum giganteum in the top center, Bryum sp. in the upper right corner, and Marchantia formosana in the lower right corner. The poster layout is clean and minimalist, with text arranged both vertically and horizontally. In

the upper right corner is the vertically arranged black Song-style characters "苔痕", below which is the horizontally arranged sans-serif English text "Moss Exhibition". On the left side is the black sans-serif text "Elkhorn Fern LifeStyle". The lower left corner reads "Alishan Moss Ecological". The date and time are prominently displayed at the bottom: "2001" in large black serif font, followed by smaller sans-serif text "04.22 [Apr.] am. 09:00" and "05.22 [May] pm. 17:00". Each plant illustration is accompanied by its scientific name, written in small gray sans-serif font, such as "Fissidens", "Rhodobryum giganteum", "Bryum sp.", "Bartramiaceae", "Alishan Moss Ecological", "Marchantia formosana", and "Astrocella yoshinagana". The overall style is elegant and educational, combining scientific illustration with modern typography.)

A.3. Figure 3

Row #1

- Case #1:
  头发变成淡紫色卷发，发丝边缘散发出金光。
  (Translation: The hair becomes light purple curls, with golden light emanating from the edges of the hair strands.)
- Case #2:
  让这个女性盘腿坐在酒店房间的地毯上，面带微笑，眼神温柔地注视着镜头，背景是整洁的白色大床和温馨的床头灯，整体氛围宁静而优雅。
  (Translation: Have this woman sit cross-legged on the hotel room carpet, with a smile on her face and a gentle gaze looking at the camera. The background features a neat white bed and warm bedside lamp, with an overall atmosphere that is tranquil and elegant.)
- Case #3:
  变成水彩风格。
  (Translation: Transform into watercolor style.)
- Case #4:
  把老鹰变成玻璃材质，同时把天空变成橙色的夕阳。
  (Translation: Transform the eagle into glass material, while changing the sky into an orange sunset.)

Row #2

- Case #1:
  参考猫的形象，生成三宫格故事，画面包括海滩、远山、落日、夕阳，三个宫格从上到下。宫格1（顶部）：猫背对着镜头，坐在沙滩上，凝望着远方的夕阳和海面，字幕"山的那边是什么"；宫格2（中部）：猫咪转过身来，侧脸对着镜头，字幕"你不必说"；宫格3（底部）：猫脸特写，正脸直视镜头，字幕"我根本不想知道"。
  (Translation: Referencing the cat's image, generate a three-panel story. The scenes include a beach, distant mountains, sunset, and evening glow, with three panels arranged from top to bottom. Panel 1 (top): The cat sits on the beach with its back to the camera, gazing at the distant sunset and sea, with subtitle "山的那边是什么"; Panel 2 (middle): The cat turns around, showing its profile to the camera, with subtitle "你不必说"; Panel 3 (bottom): Close-up of the cat's face, looking directly at the camera, with subtitle "我根本不想知道".)
- Case #2:
  制作海报，背景变成公路和蓝天白云，两侧是开阔的田野。自行车保持姿态不变，放在公路中央。海报上方是主标题"环球骑行派对"，使用粗犷、动感的黑色字体，其正下方是副标题"From 2.30 to 2.31"。
  (Translation: Create a poster with the background transformed into a highway with blue sky and white clouds, with open fields on both sides. The bicycle remains in its original posture, placed in the center of the road. At the top of the poster is the main title "环球骑行派对" in bold, dynamic black font, with the subtitle "From 2.30 to 2.31" directly below it.)
- Case #3:
  把猫换成一只有着相同姿势的哈士奇，文字"Love Cat"修改为"Love Dog"，以及文字"喵喵"改为"汪汪"。同时在左下角的宠物餐盘里加满狗粮。

(Translation: Replace the cat with a husky in the same pose, change the text "Love Cat" to "Love Dog", and change the text "喵喵" to "汪汪". Additionally, fill the pet food bowl in the lower left corner with dog food.)

Row #3

- Case #1:
  让猫和狗分别穿上粉色和和绿色的毛衣，背景改为海边的户外咖啡店，猫和狗坐在桌子旁的椅子上喝咖啡。
  (Translation: Have the cat and dog wear pink and green sweaters respectively, change the background to a seaside outdoor café, with the cat and dog sitting on chairs at the table drinking coffee.)
- Case #2:
  把这张图变成一幅画，金色边框，挂在画展的墙上，旁边有一些人在欣赏这幅画。
  (Translation: Transform this image into a painting with a golden frame, hanging on the wall of an art exhibition, with some people standing beside it appreciating the painting.)

Row #4

- Case #1:
  让一个东方美女穿上这个连衣裙和黑色的鞋，并戴上这顶帽子。美女站在公园草地上，背景有几颗桃树。
  (Translation: Have an East Asian woman wear this dress and black shoes, and put on this hat. The woman stands on the park lawn, with several peach trees in the background.)

Row #5

- Case #1:
  参考图像，生成一个狐狸玩偶的彩色图像，玩偶放在公园草地上，背景有一些树木。
  (Translation: Referencing the image, generate a color image of a fox plush toy, with the toy placed on the park lawn and some trees in the background.)
- Case #2:
  旋转玩偶，展现他的侧面。
  (Translation: Rotate the toy to show its side profile.)
- Case #3:
  生成两个这个狐狸玩偶在超市购物的画面，像人一样推着购物车购物，购物车里放满了水果。
  (Translation: Generate an image of two of these fox plush toys shopping in a supermarket, pushing shopping carts like humans, with the carts filled with fruits.)