HunyuanImage 3.0 Technical Report

Tencent Hunyuan Foundation Model Team

Abstract

We present HunyuanImage 3.0, a native multimodal model that unifies multimodal understanding and generation within an autoregressive framework, with its image generation module publicly available. The achievement of HunyuanImage 3.0 relies on several key components, including meticulous data curation, advanced architecture design, a native Chain-of-Thoughts schema, progressive model pre-training, aggressive model post-training, and an efficient infrastructure that enables large-scale training and inference. With these advancements, we successfully trained a Mixture-of-Experts (MoE) model comprising over 80 billion parameters in total, with 13 billion parameters activated per token during inference, making it the largest and most powerful open-source image generative model to date. We conducted extensive experiments and the results of automatic and human evaluation of text-image alignment and visual quality demonstrate that HunyuanImage 3.0 rivals previous state-of-the-art models. By releasing the code and weights of HunyuanImage 3.0, we aim to enable the community to explore new ideas with a state-of-the-art foundation model, fostering a dynamic and vibrant multimodal ecosystem. All open source assets are publicly available at here.

1 Introduction

In recent years, image generation models have achieved remarkable progress, enabling the generation of realistic and diverse images from natural language descriptions and reference images. Advances in deep learning architectures, particularly diffusion models [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] and transformer-based frameworks [11, 12], have significantly improved image fidelity and semantic alignment with input texts. Leading models such as Seedream 4.0 [13], Nano Banana [14] and GPT-Image [15] and HunyuanImage 2.1 [16] have demonstrated the capability to synthesize complex scenes and artistic styles or edit image accurately, attracting widespread attention in both research and industry. However, these state-of-the-art systems are predominantly closed-source, limiting transparency and reproducibility for the wider research community.

For this reason, we present HunyuanImage 3.0, an open-source model that achieves image generation performance comparable to, or surpassing, that of leading closed-source models. HunyuanImage 3.0 originates from our internally developed native multimodal model, currently with fine-tuning and post-training focused solely on image generation. We employ Hunyuan-A13B [17], a pre-trained Mixture-of-Experts (MoE) large language model (LLM) with over 80 billion total parameters, of which 13 billion are activated per token during inference, as our base model. The choice reconciles the need for both high model capacity and computational efficiency. To extend the LLM to handle visual inputs for image understanding and generation, we augment it with a pre-trained vision encoder and a VAE, each equipped with a projection layer that transforms the extracted image features into a joint embedding space compatible with the LLM's word embeddings [18, 19, 20, 21, 22]. For image understanding, the LLM conditions its autoregressive next-token prediction on joint image features extracted from the vision encoder and the VAE to generate appropriate responses. For image generation, diffusion-based image modeling on VAE image features is incorporated into the LLM in the same manner as in Transfusion [21] and JanusFlow [22]. Furthermore, the LLM-based framework enables us to incorporate Chain-of-Thought training and inference, thereby improving the

Figure 1: Multi-ratio text-to-image samples from HunyuanImage 3.0, demonstrating its powerful prompt-following, reasoning, concept generalization and text-rendering capabilities.

performance of both image understanding and generation tasks. After fine-tuning and post-training the pre-trained model solely on image generation tasks, we establish the image generation module of HunyuanImage 3.0, which currently stands as the largest and most powerful open-source image generation model. We conduct extensive experiments on both automatic and human evaluation, the results on text-image alignment and visual quality demonstrate that HunyuanImage 3.0 rivals previous state-of-the-art models, including Seedream 4.0 [13], Nano Banana [14], GPT-Image [15] and HunyuanImage 2.1 [16]. By releasing the code and weights of HunyuanImage 3.0, we aim to enable the community to explore new ideas with a state-of-the-art foundation model, fostering a dynamic and vibrant image generation ecosystem.

This report is structured as follows. In **Section 2**, we introduce our data preparation techniques, including filtering and captioning models. **Section 3** presents detailed information about the architecture and algorithms of HunyuanImage 3.0. In **Section 4**, we discuss our training strategies and algorithms. In **Section 5**, we evaluate the performance of HunyuanImage 3.0 and compare it with state-of-the-art text-to-image generation models.

2 Data Preparation

2.1 Data Filtering

To construct a diverse, high-quality image dataset, we implemented a comprehensive three-stage filtering process on an initial pool of over 10 billion raw images. This rigorous process, which ultimately retained less than 45% of the initial data, was designed to prioritize both semantic diversity and image quality, critical requirements for training robust generative models.

In the first stage, we addressed technical flaws by removing images with low resolution (less than 512 pixels), broken files, over-/under-exposure, and over-saturation. We also deduplicated the images according to their MD5 values.

The second stage served as our primary data curation process, employing two types of operators: objective filters and subject-scoring operators. The objective filters were learning-based detectors for watermarks, logos, extensive text (through the hy-OCR model[1]), collages, prominent borders and AI-generated content (AIGC). To maintain accuracy when dealing with massive data, these detectors were trained using balanced training datasets created through stratified sampling. The proliferation of AIGC images poses a significant challenge by distorting natural data distributions and impairing model convergence. Our mitigation strategy combined an automated AIGC detection model [23, 24, 25] with the removal of all images from data sources found to have a high proportion of AI-generated content.

Our subject-scoring operators included models for image clarity and aesthetics. The clarity model provided an overall score based on an image's sharpness, noise level, and dynamic range. To ensure consistent and interpretable aesthetic scores, our artists systematically designed a criterion that considered three fundamental elements: color, light & shadow, and composition. Based on this criterion, we build our own aesthetics model. We applied a unified threshold value to filter out unqualified images across all types, while using different threshold values for specific genres to select data for later training stages.

In the final stage, we further deduplicated data based on embedding cluster results, which removed approximately 0.5% of the data, making our datasets more compact. To enhance semantic breadth, we strategically supplemented the filtered set with specialized datasets, including knowledge-augmented , text-related, stylized, and graphic design collections. This meticulous pipeline resulted in clean, high-quality, and diverse datasets formed with nearly 5 billion images for training advanced generative models. Beyond the single-image corpus, we constructed a specialized dataset of over 100 million image pairs and multi-image samples specifically designed for learning interleave relationships. The image-pair subset was derived through two primary approaches: image clustering and video segment mining.

For the image clustering approach, following the analysis of over two billion image clusters, we selectively extracted pairs from the representative clusters that exhibited potential similarity. These pairs were then passed through an image-relation discrimination operator to retain only those with

[1] to be released

a significant editing relationship. To optimize the model's learning efficacy, an image complexity model [26] was applied to filter out images whose constituent elements were deemed overly complex.

The video data mining pipeline began with shot boundary detection to isolate video segments corresponding to unified scenes. Camera motion classification operator was subsequently employed to exclude clips exhibiting excessive camera transformation. We further refined the selection by integrating results from object detection and semantic segmentation to isolate keyframes demonstrating canonical transformation relationships. Finally, to mitigate the detrimental effects of motion blur on model training, the selected frames underwent an additional round of filtering based on a motion blur detection operator. The resulting multi-image data is composed of interleaved data (image-text sequences) sourced from the internet and extracted frames from videos.

2.2 Image Captioning

To generate rich, controllable, and factually-grounded image descriptions, we propose a novel pipeline built upon three core components: (1) a hierarchical schema for structured image description, (2) a compositional synthesis strategy for diverse data augmentation, and (3) specialized agents for factual grounding, as illustrated in Figure 2 .

Bilingual and Hierarchical Captioning Schema. The foundation of our approach is a bilingual (English/Chinese) and hierarchical captioning schema that decomposes image content into multiple, well-defined semantic fields. This structured representation includes:

- **Descriptive Levels (Short to Extra-Long)**: Four tiers of narrative detail, from a concise summary to an exhaustive depiction of all foreground and background elements.
- **Stylistic Attributes**: Fields for capturing the image's artistic style, cinematographic shot type, lighting, prevailing atmosphere, and composition.
- **Factual Entities**: A dedicated field for named entities (IP), identifying specific characters, landmarks, brands, and artworks.

This hierarchical schema not only enables fine-grained control over the generative process, but also serves as the structural basis for our data synthesis engine.

Compositional Caption Synthesis for Data Diversity. To enhance model generalization and mitigate overfitting, we introduce Compositional Caption Synthesis, a dynamic data augmentation strategy that leverages our hierarchical schema. During training, we strategically sample and combine different fields to generate captions varying in both length and pattern, supporting bilingual (English/Chinese) outputs from about 30 words up to 1,000 words.

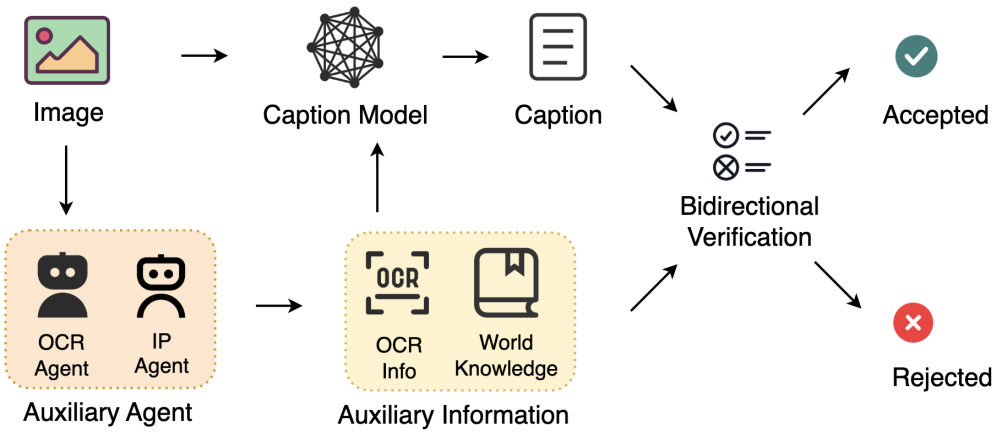

Figure 2: Image Captioning Pipeline.

Factual Grounding via Specialized Agents and Bidirectional Verification. To overcome the limitations of standard Visual Language Models (VLMs) in recognizing dense in-image text and entities requiring world knowledge, we integrate two specialized agents to ground the descriptions in

verifiable reality. An OCR Agent extracts in-image text, while a Named Entity (IP) Agent identifies real-world entities. This external knowledge is fed as auxiliary input to the caption model. Crucially, we establish a Bidirectional Verification Loop that cross-references the entities detected by the agents with the generated caption. Following this, only the samples that successfully pass the bidirectional verification are included in the final training dataset.

Image Difference Captioning. Additionally, for paired image data, we have developed an image difference captioner. This model takes a pair of images, their captions and corresponding two-frame video as input to generate a caption detailing the changes in the foreground and background, serving to simulate user-input editing instruction. The contextual information is important to enable the model to generate more accurate and detailed difference descriptions.

2.3 Reasoning Dataset Construction

Our model is a powerful, natively multi-modal architecture endowed with robust reasoning and semantic understanding capabilities. A key contribution of our work is the elicitation of an automated *Chain-of-Thought (CoT)* reasoning process for image generation, which is activated through fine-tuning on a small, specialized dataset. This process enables the model to autonomously execute a full pipeline: from interpreting an initial input prompt, to engaging in an intermediate "thinking" phase of conceptual refinement and rewriting, and finally to synthesizing the target image. To effectively unlock this latent ability, we constructed two specific types of training data: **(1)** Text-to-Text (**T2T**) reasoning data to bolster its logical inference. **(2)** Text-to-Text-to-Image (**T2TI**) reasoning data that explicitly models the entire workflow from abstract concepts to their visual manifestations. This training strategy empowers the model to achieve a seamless, automated, and coherent translation from high-level user intent to high-fidelity visual output.

Text to Text. To enhance the mdoel's instruction following and logical-reasoning ability, we curated a diverse textual corpus of real-world image generation prompts. This corpus spans photorealistic rendering, artistic and stylistic renderings, UI and poster design tasks, knowledge-driven queries, and scientific or technical visualizations. By covering a broad spectrum of user intents, domains, and complexity levels, the model trained with T2T data can parse nuanced requirements, resolve ambiguities, and produce coherent, stepwise textual reasoning that faithfully maps instructions to precise image captions.

Text to Text&Image. To improve end-to-end textual reasoning and visual fidelity, we developed the T2TI corpus, a high-quality, class-balanced image dataset. Images were filtered from the the pretraining dataset using aesthetic metrics and paired with their original short and long captions. We also compiled a collection of infographics from Wikipedia. For each image, we annotate a corresponding reasoning trace that refines goals and translates user intent into detailed visual specification. The images along with their captions, reasoning traces are used to improve the model's CoT image generation ability.

3 Model Design

3.1 Native Multimodal Model

We present a native multimodal model designed for unified understanding and generation across text and image modalities. As illustrated in Figure 3, the proposed architecture adopts a hybrid discrete-continuous modeling strategy: text tokens are modeled via autoregressive next-token prediction, while image tokens are modeled through a diffusion-based prediction framework [7]. The system, designated as HunyuanImage 3.0, brings together language modeling, image understanding, and image generation in a cohesive framework to achieve unified multimodal modeling.

3.1.1 Architecture

Backbone. The model backbone is constructed upon the Hunyuan-A13B [17], a decoder-only LLM over 80 billion total parameters. It employs an MoE configuration comprising 64 experts, with 8 experts activated per token, accompanied by one shared multi-layer perceptron (MLP). This design results in approximately 13 billion activated parameters per token, balancing expressive capacity with computational efficiency.

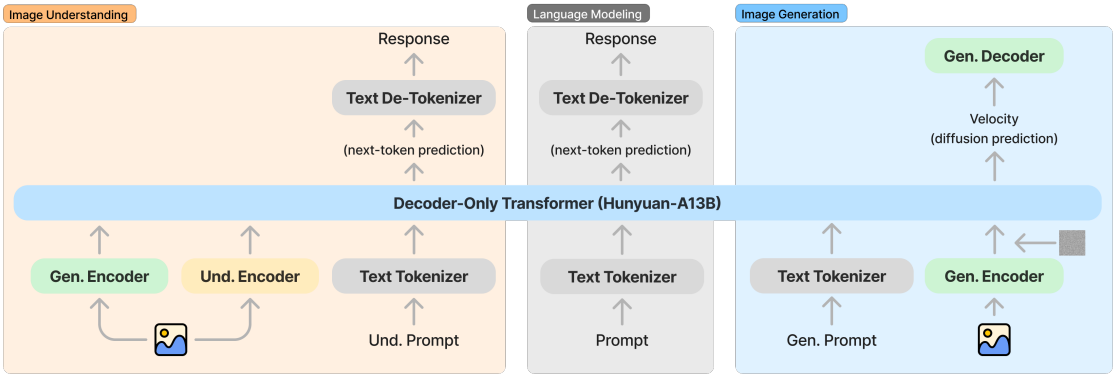

Figure 3: Illustration of HunyuanImage 3.0.

Text Tokenizer. For textual input, we utilize the Hunyuan Tokenizer [17], extending its vocabulary with a set of custom special tokens tailored to support image generation and understanding tasks.

Image Encoder. In the image generation pathway, we employ an internal VAE that projects raw pixel values into a 32-dimensional latent space with a downsampling factor of 16. Prior approaches, such as those in DiT-like [27] architectures [28, 29, 30, 31] typically combined an 8x downsampling VAE with an additional patchification layer that further reduced spatial resolution by a factor of 2. In contrast, we demonstrate that a single VAE with 16x downsampling offers a simpler and more effective alternative, yielding superior image generation quality.

For conditioned image inputs, we introduce a dual-encoder strategy that concatenates latent features from the VAE with those from a vision encoder. This approach enables unified multimodal representation that supports both generation and understanding within a single sequence—a key different from previous unified models [32, 33, 34, 35], which often segregated visual features by task (e.g., using vision encoder features for understanding and VAE features for generation). Our method facilitates complex multimodal interactions—such as interleaved text dialogue, image generation, image understanding, and image-editing—within a continuous context, thereby eliminaing the need to switch between separate understanding and generation pipelines.

Projector. We design two distinct projectors modules to align features from the dual image encoders into the transformer's latent space. Features from the VAE are projected using a timestep-modulated residual block [27, 36], whereas features from the vision encoder are transformed via a two-layer MLP. Additionally, we incorporate timestep embedding into the sequence to enhance the conditioning of the diffusion process.

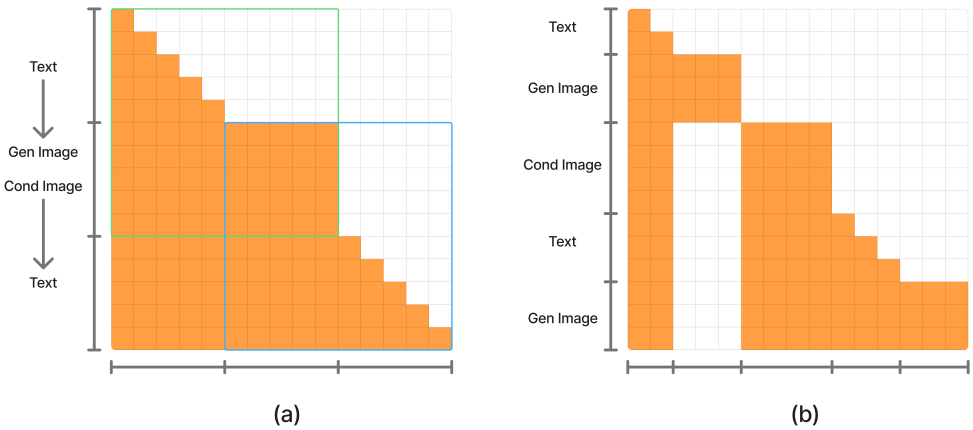

Figure 4: Two types of attention implementation.

3.1.2 Generalized Causal Attention

Causal attention is a fundamental component in LLMs for autoregressive text generation, as it ensures each token only attends to preceding tokens, thereby preserving the autoregressive property. In contrast, full attention is commonly employed in DiTs for image generation, allowing each image token to attention to all other tokens within the same image, which is beneficial for capturing global spatial dependencies. In our proposed native multimodal model, we integrate both attention types to handle heterogeneous data modalities effectively. Specifically, we introduce a *Generalized Causal Attention* mechanism. Within this scheme, text tokens are restricted to attend only to previous multimodal tokens in the sequence. Image tokens, however, are permitted to attend to all previous multimodal tokens as well as all successive image tokens within the same image segment. This design respects the autoregressive geneartion nature of text while leveraging the global contextual capacity of full attention for image patches.

As illustrated in Figure 4 we categorize the training attention mask into two distinct types based on the number of generated image segments (Gen Image), which correspond to the noised images being processed. In sequences where there are no Gen Images (as in image understanding tasks, indicated by the blue box in Figure 4 (a)) or exactly one Gen Image (as in text-to-image tasks, green box in Figure 4 (a)), the attention mask adheres strictly to the Generalized Causal Attention pattern defined above. However, when multiple Gen Images are present within a single training sequence (Figure 4 (b)), a modification is necessary: any Gen Images appearing in the context must not be attended to by subsequent tokens in the sequence. This constraint introduces a "hole" (i.e., a region of mask attention) in the lower triangular part of the attention mask.

During inference, the input sequence never contains more than one simultaneous Gen Image. This is because once an image is generated, it is treated as a conditional image (Cond Image) for subsequent tokens in the sequence. Thus, the attention mask during inference consistently follows the cananical Generalized Causal Attention structure without requiring the additional masking needed during multi-gen-image training. This approach ensures causal consistency in generation while enabling effective multimodal learning.

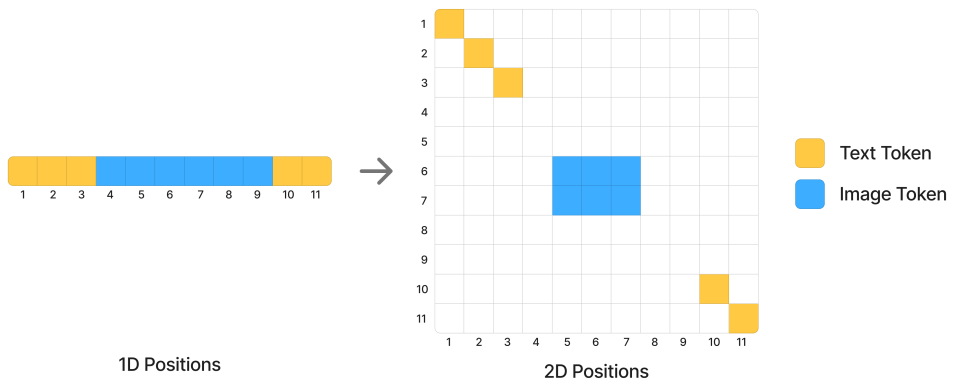

Figure 5: Illustration of the comparison between 1D RoPE and *Generalized 2D RoPE* with backward compatibility. The text tokens in 1D RoPE has exactly the same positions as 2D RoPE. Intuitively, this generalization is implemented by reshape the 1D image positions into 2D positions and place it in the middle of two text sections.

3.1.3 Position Embedding

Rotary Position Embedding (RoPE) [37] is widely adopted in LLMs and DiTs due to its flexibility and scalability. In this work, we implement a *Generalized 2D RoPE*, as proposed by Su. [2] This approach maintains backward compatibility with the pretrained LLM. Formally, for a one-dimensional text position index n, and a set of frequencies $\{\theta_0, \theta_1, \dots\}$, the position embedding is defined as $[\cos(n\theta_0), \cos(n\theta_1), \dots, \sin(n\theta_0), \sin(n\theta_1), \dots]$. We generalize this formulation to

[2] https://kexue.fm/archives/10352

two-dimensional coordinates by interpreting the same embedding structure anisotrapically. For a position (x, y), the embedding becomes $[\cos(x\theta_0), \cos(y\theta_1), \ldots, \sin(x\theta_0), \sin(y\theta_1), \ldots]$. As depicted in Figure 5, image tokens—which are reshaped from 1D to 2D—are assigned such generalized 2D position encodings, while text tokens retain standard 1D RoPE, and also can be viewed as 2D RoPE of diagnoal positions. This design ensures that in the absence of image tokens, the encoding reduces exactly to 1D RoPE, thereby preserving full compatibility with conventional text generation and minimizing disruptive effects on pre-trained linguistic capabilities.

In the training sequence incorporating multiple Gen Images (Figure 4 (b)), the tokens following each Gen Image are assigned different positions in the training and inference sequences. To ensure positional consistency between training and inference, the position embeddings for these tokens are adjusted by shifting their token positions accordingly. This alignment is critical for maintaining the structural integrity of the sequence during model training and inference, as it mitigates potential discrepancies introduced by the variable placement of generated images. By explicitly aligning the positional encoding scheme across both phases, the model can more effectively generalize and preserve coherent contextual relationships within the sequence.

3.1.4 Automatic Resolution

DiT-like models typically require deterministic user input to specify the desired image size and aspect ratio. In the proposed native multimodal model, we introduce an automatic mode that allows the model to determine appropriate image shapes based on the context, which can be use prompt or conditional image tokens. Specifically, we extend the language model's vocabulary with two groups of special tokens: one set represented as {<img_size_256>, <img_size_512>, <img_size_768>, ...}, and the other as {<img_ratio_0>, <img_ratio_1>, <img_ratio_2>, ..., <img_ratio_32>}. Each <img_size_*> token corresponds to an image resolution anchor, while each <img_ratio_*> token represents an aspect ratio ranging from 1:4 to 4:1. During training, the model learns to associate these shape tokens with the user inputs and previous conversations in the context, enabling it to predict appropriate size and ratio tokens according to the input context. Additionally, users can provide explicit cues—such as "3:4" or "vertical"—to guide the model toward generating a specific aspect ratio token. Based on the predicted size and ratio tokens, we can incorporate the 2D RoPE for the image tokens, enabling the model to generate images with the desired structural properties.

4 Model Training

4.1 Pre-training

Given the diversity of tasks involving heterogeneous token sequences in multimodal modeling, we design a flexible multi-task training framework capable of supporting large-scale training across numerous tasks and datasets, including text-to-image generation (T2I), language modeling (LM), multimodal understanding (MMU), interleaved text-image modeling (INTL) and reasoning (CoT). Our pre-training process is organized into four progressive stages, in which training data are filtered from coarse to fine and image resolutions are increased gradually for VAE encoder and fixed for ViT encoder. During training, the aspect ratio of images are preserved to enable the multi-resolution image generation. Detailed configurations for each stage are provided in Table 1.

Table 1: Training stages of the proposed native multimodal model. Resolution anchor denotes that the images are resized to the desired size while keeping the aspect ratio. We adopt progressive image resolution anchors for VAE and a fixed resolution anchor for vision encoder (ViT).

Training Stage	VAE Reso. Anchor	ViT Reso. Anchor	Training Part	Task
I	256px	512px	Transformer	T2I, LM, MMU
II	256px	512px	ViT	MMU
III	512px	512px	ViT, Transformer	T2I, LM, MMU, INTL
IV	1024px	512px	ViT, Transformer	T2I, LM, MMU, INTL, CoT

Progressive Training. In the first stage, we train the Transformer backbone while keeping the ViT frozen. Three tasks are optimized simultaneously: text-to-image (T2I), language modeling (LM), and multimodal understanding (MMU), utilizing both text-image pairs and text-only data.

This stage employs a low image resolution (256px) for VAE encoder and a large batch size training, enabling the model to learn from billions of images and align the latent representations of text and image modalities. During the second stage, the Transformer backbone remains frozen, while the ViT and its associated aligner module are fine-tuned using only MMU data to enhance visual understanding capabilities. In Stage III, both the ViT and Transformer are jointly trained with images of higher resolution (greater than 512px). The dataset size is reduced to increase the proportion of high-quality images. Interleaved text-image data, such as image editing and image-to-image data, are incorporated at this stage to enhance multimodal modeling capabilities. In the final stage, training images are further constrained to a high-resolution subset, each with at least 1024 pixels on the shorter edge. Similarly, images used for the MMU task are limited to a high-resolution subset to enhance understanding capability. Although the input image size for the ViT encoder remains fixed at 512 pixels, we observe that high-resolution VAE features also contribute to improved model understanding. In addition, reasoning data introduced in Section 2.3 are incorporated at this stage to enable reasoning in multimodal modeling, particularly for Chain-of-Thoughts-based text-to-image generation. Significantly, tokens of reasoning part are also modeled via autoregressive next-token prediction.

Instruction Tuning. After pre-training our native multimodal large language model, we perform instruction tuning specifically tailored for text-to-image generation. At this stage, T2I, LM, and CoT data are formatted using instruction templates and jointly used to optimize the model.

4.2 Post-training

The post-training optimization of our model is a multi-stage process designed to systematically refine its generative capabilities. We first conduct SFT on a meticulously curated dataset of human-annotated examples. Following this, DPO [38] is implemented to effectively address and reduce physical distortions. We then utilize MixGRPO [39] to enhance the critical aspects of text-image alignment, realism, and aesthetic appeal. The final refinement is achieved through the application of SRPO [40] and a novel, in-house Reward Distribution Alignment (ReDA) method, which together are instrumental in further elevating the realism and clarity of the generated images.

SFT. For the SFT stage, we collect carefully curated high quality images encompassing diverse categories, including landscapes, portraits, animals, OCR, and others. We implement a multi-stage training strategy where subsequent stages incorporate progressively higher-quality training samples.

DPO. We utilize DPO to address common issues of structural deficiency in image generation. The training data is prepared by first generating a corpus of images from the SFT model. These images are subsequently annotated to create a paired dataset of high-quality and low-quality samples. This dataset serves as a preference signal, which is applied to effectively suppress distortions and improve visual appeal.

MixGRPO. MixGRPO is an efficient online reinforcement learning framework that extends GRPO to flow-based models through a hybrid ODE–SDE sampling strategy. We apply MixGRPO with both open-source and proprietary reward models to optimize aesthetics (style, composition, lighting), mitigate distortions, and reduce artifacts. In addition, we refine advantage estimation to accelerate convergence, and demonstrate the scalability of MixGRPO to large-scale training regimes, achieving stronger alignment with human preferences.

SRPO. SRPO is a novel gradient-guided online reinforcement training strategy designed to enhance the realism and aesthetic quality of generated images. It directly injects a noise prior into the latent space features and then denoises it to a clean image in a single step. It selects the initial interval of the denoising trajectory for optimization, where the model has greater flexibility for improvement. By incorporating differentiable reward signals from both positive and negative text guidance, the model can efficiently align with human preferences and mitigate common issues in AI-generated images, such as oversaturation, incoherent lighting and colors, and poor skin texture.

ReDA. We've developed a novel reward distritution alignment algorithm, ReDA, for the post-training phase of our generative model. ReDA effectively improves visual quality by minimizing the

divergence between the model's generated outputs and a high-reward distribution, which is defined by a diverse set of high-quality images from various genres.

5 Model Performance

5.1 SSAE

Modern Text-to-Image (T2I) generative models rely on standardized benchmarks like T2I-CompBench [41] and GenEval [42] to measure progress. However, these benchmarks exhibit limitations in comprehensiveness and reliability:

(1) Deficiencies in Prompt Design and Semantic Diversity: They often use short, formulaic structures (e.g., "a photo of a [object] with [attribute]"), failing to capture the complexity of real-world user instructions. This lack of diversity inadequately stress-tests model capabilities in parsing longer, intricate natural language descriptions involving multi-attribute composition, complex relational reasoning, and contextual logic.

(2) Over-reliance on Automated Metrics Misaligned with Human Judgment: Both benchmarks depend heavily on automatic metrics like CLIP Score for evaluating text-image alignment. However, these metrics are poor proxies for human assessment, as they may highly rate images with critical failures in spatial relationships (e.g., confusing "a boy under a bee" with "a bee under a boy") or precise attribute binding. This creates a disconnect between benchmark scores and human-perceived utility.

To address these issues, we propose a **s**tructured **s**emantic **a**lignment **e**valuation metric, *abbr.*, **SSAE**. This intelligent metric leverages advanced LLMs and MLLMs for image-text alignment.

To resolve (1), we collect 500 diverse prompts and extract 3,500 key points using an LLM-based structured semantic point parser. Through in-context learning, points are categorized into 12 fine-grained fields: Nouns, main attributes and actions of primary and secondary subjects, other attributes of primary subjects, nouns and attributes of the scene, as well as camera shot, style, and composition. Another LLM then examines coherence between extracted points and original prompts, filters hallucinated points (e.g., objects or relationships inconsistent with prompts), and complements missing points, followed by human rectification. These points remain fixed during subsequent MLLM assessment to ensure stable and fair comparison across T2I models.

To resolve (2), an advanced MLLM with Chain-of-Thought reasoning scores model-generated images based on the prompts and pre-extracted key points, performing 0-1 matching. From this, we calculate both field-specific accuracy and two overall metrics: Mean Image Accuracy (mean accuracy of image-wise averaged scores) and Global Accuracy (averaged score across all key points in the dataset).

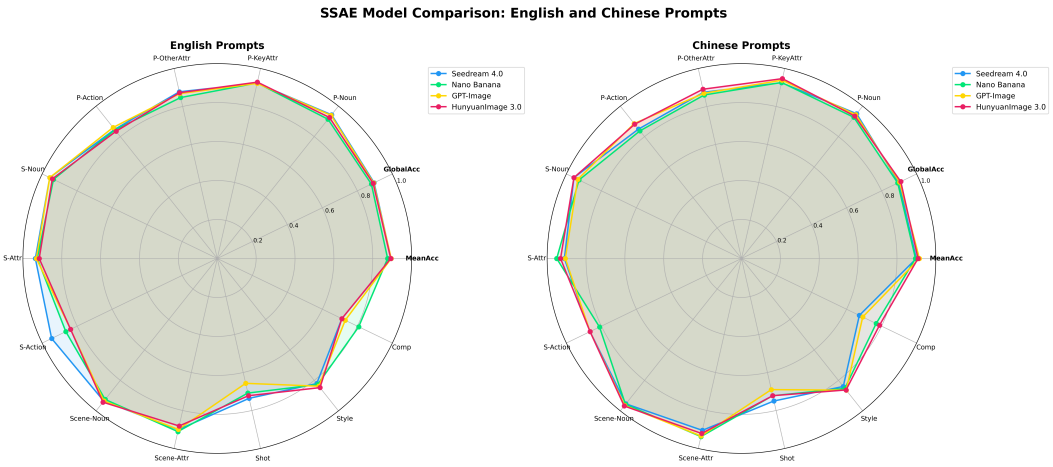

Figure 6: SSAE evaluation results.

Compared to recent benchmarks like DreamBench++ [43], which also uses MLLMs for human-like assessment, our benchmark features a more robust and comprehensive hierarchical semantic

point parsing paradigm, taxonomy, and richer extensions (e.g., ranking mode beyond scoring). The SSAE evaluation results for our method and competitors are shown in Figure 6. As illustrated, HunyuanImage 3.0 achieves performance on par with leading models in all fine-grained fields.

5.2 GSB

We adopt the GSB (Good/Same/Bad) evaluation method, which is commonly used to assess the relative performance of two models from an overall image perception perspective. In practice, we carefully construct 1,000 text prompts to cover a blanced scenarios, and generate an equal number of image samples for each model in a single run. For fairness, inference is performed only once for each prompt, without any cherry-picking of results. All other models are evaluated under their default settings. The evaluation is conducted by over 100 professional evaluators.

Figure 7 presents the GSB evaluation results. As shown, HunyuanImage 3.0 achieves a relative win rate of 14.10% compared to HunyuanImage 2.1, which was previously the best open-source model, thereby establishing HunyuanImage 3.0 as the most powerful open-source text-to-image model to date. Moreover, HunyuanImage 3.0 achieves relative win rates of 1.17%, 2.64%, and 5.00% compared to Seedream 4.0, Nano Banana, and GPT-Image, respectively. These results demonstrate that HunyuanImage 3.0, as an open-source model, has reached a level of image generation quality comparable to leading closed-source commercial models.

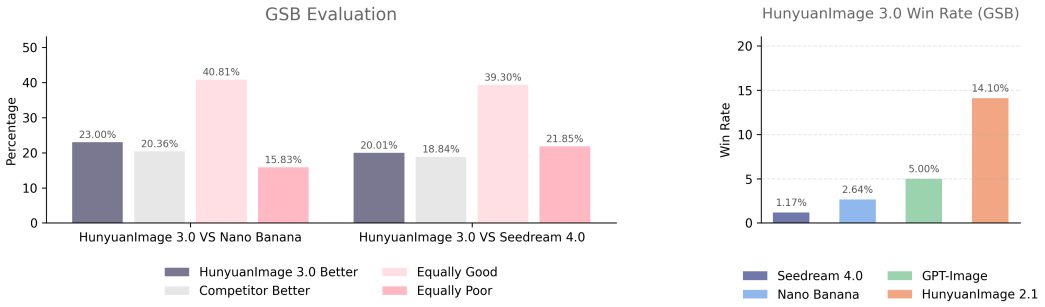

Figure 7: GSB evaluation results.

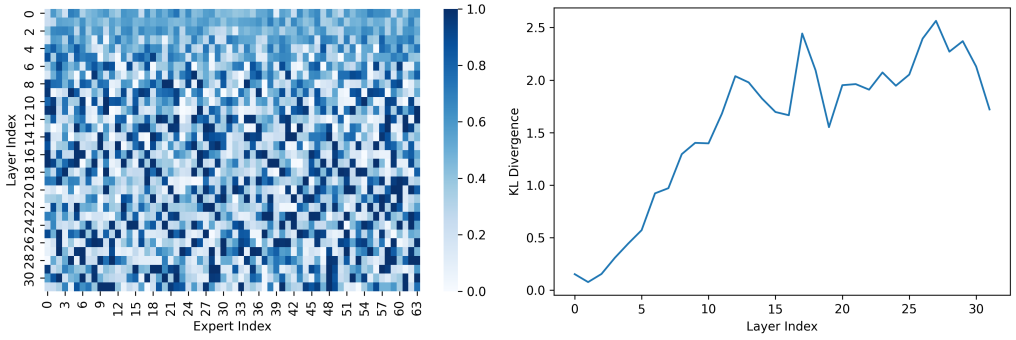

Figure 8: Left: Heatmap of $\frac{\frac{v_{ij}}{\sum_j v_{ij}}}{\frac{v_{ij}}{\sum_j v_{ij}}+\frac{t_{ij}}{\sum_j t_{ij}}}$, where v_{ij} and t_{ij} denote the image-token and text-token activation counts, respectively, for the j-th expert in the i-th layer. The darker an expert appears, the more specialized it is to image tokens. Right: KL divergence between $\{\frac{v_{ik}}{\sum_j v_{ij}}\}_{k=0}^{64}$ and $\{\frac{t_{ik}}{\sum_j t_{ij}}\}_{k=0}^{64}$ for all experts of each layer. As the layer goes deeper, the KL divergence increases and the expert activation distributions become more dispersed across modalities.

5.3 Discovery

5.3.1 Expert Activation Analysis

We analyse how experts are activated by tokens of different modals in multimodal MoE model. We ramdomly select 1,000 prompts to perform text-to-image generation and conduct statistical analysis on experts of each layer using our pre-trained model. Figure 8 demonstrates an expert modal preference heatmap and a KL divergence tendency between image- and text-activated-expert distribution for experts of each layer. Both figures imply that the experts become increasingly specialized in individual modalities. This suggests that MoE may enhance multimodal modeling by dispersing responsibilities for different modalities among specialized experts.

6 Conclusion

In this report, we present HunyuanImage 3.0, a native multimodal model that unifies multimodal understanding and generation within an autoregressive framework. We begin with a pre-trained MoE LLM and extend it to support both image understanding and generation. With large-scale pre-training on diverse and carefully curated multimodal data, our model demonstrates robust capabilities in both image understanding and generation. Thanks to the LLM-based framework, we incorporate native Chain-of-Thought training and inference, which improves the multimodal performance significantly. Building upon the pre-trained model, we perform fine-tuning and post-training specifically for image generation, and make the resulting model publicly available. HunyuanImage 3.0 exhibits strong capabilities in prompt-following, reasoning, concept generalization, and text rendering for text-to-image generation. Results from both automatic and human evaluations on text-image alignment and visual quality indicate that HunyuanImage 3.0 rivals existing state-of-the-art models. While this release only includes the text-to-image ability, training for image-to-image tasks is ongoing, and this capability will be released in the near future.

7 Project Contributors

- **Project Sponsors:** Jie Jiang, Linus, Peng Chen, Yuhong Liu
- **Project Leaders:** Zhao Zhong
- **Core Contributors (First project leader, otherwise listed alphabetically):**
  - **Captioner & Data:** Xin Li, Duojun Huang, Xinchi Deng, Xuefei Zhe
  - **VAE & Model Accleration:** Songtao Liu, Changlin Li, Jianbing Wu, Yang Li, Peizhen Zhang
  - **Algorithm & Pretraining:** Miles Yang, Fanbin Lu, Jian-Wei Zhang, Shi-Xue Zhang, Zijian Zhang
  - **Post Training:** Lucas Wang, Chunyu Wang, Hangting Chen, Hao Wen, Junzhe Li, Lucaz Liu, Xiangwei Shen, Yingfang Zhang, Ying Dong, Yixuan Shi, Yutao Cui, Zheng Yuan, Zhengkai Jiang, Zhimin Li
- **Contributors (Listed alphabetically):** Siyu Cao, Yiji Cheng, Kipper Gong, Tianpeng Gu, Xiusen Gu, Tiankai Hang, Weijie Kong, Donghao Li, Jiaxin Lin, Yanxin Long, Shu Liu, Yu Liu, Qinglin Lu, Yuanbo Peng, Qi Tian, Jiale Tao, Yangyu Tao, Pengfei Wan, Kai Wang, Lei Wang, Linqing Wang, Qixun Wang, Weiyan Wang, Yue Wu, Senhao Xie, Fang Yang, Xiaofeng Yang, Xuan Yang, Zhantao Yang, Jingmiao Yu, Chao Zhang, Yepeng Zhang, Zihao Zhang, Zhiyuan Zhao, Penghao Zhao, Jianchen Zhu

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[4] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[5] 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.

[6] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

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

[8] 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.

[9] Zijian Zhang, Zhou Zhao, and Zhijie Lin. Unsupervised representation learning from pre-trained diffusion probabilistic models. *Advances in neural information processing systems*, 35: 22117–22130, 2022.

[10] Zijian Zhang, Zhou Zhao, Jun Yu, and Qi Tian. Shiftddpms: Exploring conditional diffusion models by shifting diffusion trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3552–3560, 2023.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[12] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[13] ByteDance. Seedream 4.0, 2025. URL `https://seed.bytedance.com/en/seedream4_0`.

[14] Google. Nano banana, 2025. URL `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image`.

[15] OpenAI. Gpt-image, 2025. URL `https://platform.openai.com/docs/models/gpt-image-1`.

[16] Tencent Hunyuan. Hunyuanimage 2.1, 2025. URL `https://github.com/Tencent-Hunyuan/HunyuanImage-2.1`.

[17] Tencent Hunyuan Team. Hunyuan-a13b. `https://github.com/Tencent-Hunyuan/Hunyuan-A13B`, 2024.

[18] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022.

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[21] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

[22] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751, 2025.

[23] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2024.

[24] Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. *arXiv preprint arXiv:2507.02664*, 2025.

[25] Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025.

[26] Tinglei Feng, Yingjie Zhai, Jufeng Yang, Jie Liang, Deng-Ping Fan, Jing Zhang, Ling Shao, and Dacheng Tao. Ic9600: A benchmark dataset for automatic image complexity assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (01):1–17, 2023. doi: 10.1109/TPAMI.2022.3232328.

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[28] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. URL `https://openreview.net/forum?id=FPnUhsQJ5B`.

[29] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024. URL `https://arxiv.org/abs/2405.08748`.

[30] Black Forest Labs. Flux, 2024. URL `https://blackforestlabs.ai/`.

[31] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL `https://arxiv.org/abs/2412.03603`.

[32] 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. *arXiv preprint arXiv:2408.12528*, 2024.

[33] 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.

[34] 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. *arXiv preprint arXiv:2505.14683*, 2025.

[35] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[37] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[38] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[39] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

[40] Xiangwei Shen, Zhimin Li, Zhantao Yang, Shiyi Zhang, Yingfang Zhang, Donghao Li, Chunyu Wang, Qinglin Lu, and Yansong Tang. Directly aligning the full diffusion trajectory with fine-grained human preference. *arXiv preprint arXiv:2509.06942*, 2025.

[41] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[42] 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.

[43] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024.