

LUMINA-IMAGE 2.0: A UNIFIED AND EFFICIENT IMAGE GENERATIVE FRAMEWORK

Qi Qin^{2*}, Le Zhuo^{1*}, Yi Xin^{1,4*}, Ruoyi Du^{1*}, Zhen Li^{3*}, Bin Fu^{1*}, Yiting Lu^{1*}, Jiakang Yuan¹, Xinyue Li¹, Dongyang Liu^{1,3}, Xiangyang Zhu¹, Manyuan Zhang³, Will Beddow⁶, Erwann Millon⁶, Victor Perez⁶, Wenhui Wang¹, Conghui He¹, Bo Zhang¹, Xiaohong Liu⁵, Hongsheng Li³, Yu Qiao¹, Chang Xu², Peng Gao^{1†‡}

¹ Shanghai AI Laboratory, ² The University of Sydney, ³ The Chinese University of Hong Kong,

⁴ Shanghai Innovation Institute, ⁵ Shanghai Jiao Tong University, ⁶ Krea AI

ABSTRACT

We introduce **Lumina-Image 2.0**, an advanced text-to-image generation framework that achieves significant progress compared to previous work, Lumina-Next. Lumina-Image 2.0 is built upon two key principles: (1) *Unification* – it adopts a unified architecture (Unified Next-DiT) that treats text and image tokens as a joint sequence, enabling natural cross-modal interactions and allowing seamless task expansion. Besides, since high-quality captioners can provide semantically well-aligned text-image training pairs, we introduce a unified captioning system, Unified Captioner (UniCap), specifically designed for T2I generation tasks. UniCap excels at generating comprehensive and accurate captions, accelerating convergence and enhancing prompt adherence. (2) *Efficiency* – to improve the efficiency of our proposed model, we develop multi-stage progressive training strategies and introduce inference acceleration techniques without compromising image quality. Extensive evaluations on academic benchmarks and public text-to-image arenas show that Lumina-Image 2.0 delivers strong performances even with only 2.6B parameters, highlighting its scalability and design efficiency. We have released our training details, code, and models at <https://github.com/Alpha-VLLM/Lumina-Image-2.0>.

1 Introduction

Text-to-image (T2I) generative models have made significant strides over the past years. Notable open-source models [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] have shown significant improvements in both image fidelity and prompt adherence, thereby broadening the scope of their applicability in diverse downstream tasks [11, 12, 13, 14]. From these advancements, (1) the scalable text-conditional Diffusion Transformer (DiT) architectures, and (2) the large-scale, high-quality text-image datasets are witnessed as the most important factors for developing text-conditional image generative models.

However, existing models still exhibit notable limitations in both aspects. First, many text-conditional Diffusion Transformers [15, 8, 16, 17, 18, 19] continue to rely on cross-attention mechanisms to inject textual information. This paradigm treats text embeddings as fixed external features, thus limiting the efficiency of multimodal fusion and may even introduce uni-directional bias when using text embedding extracted from causal large language models [20]. Moreover, extending these models to new tasks often requires specific architecture designs [11, 21]. Second, although recent efforts [2, 15, 22] have highlighted the importance of collecting high-quality image captions, the lack of a dedicated captioning system tailored for T2I generation has resulted in inaccurate and insufficiently image captions for text-image paired training data. The limitations in both architecture and data quality constrain the expressiveness of text and visual representations, ultimately impairing the model’s ability to faithfully follow user instructions in generating high-quality images.

*Equal Contribution

†Corresponding Authors

‡Project Leader

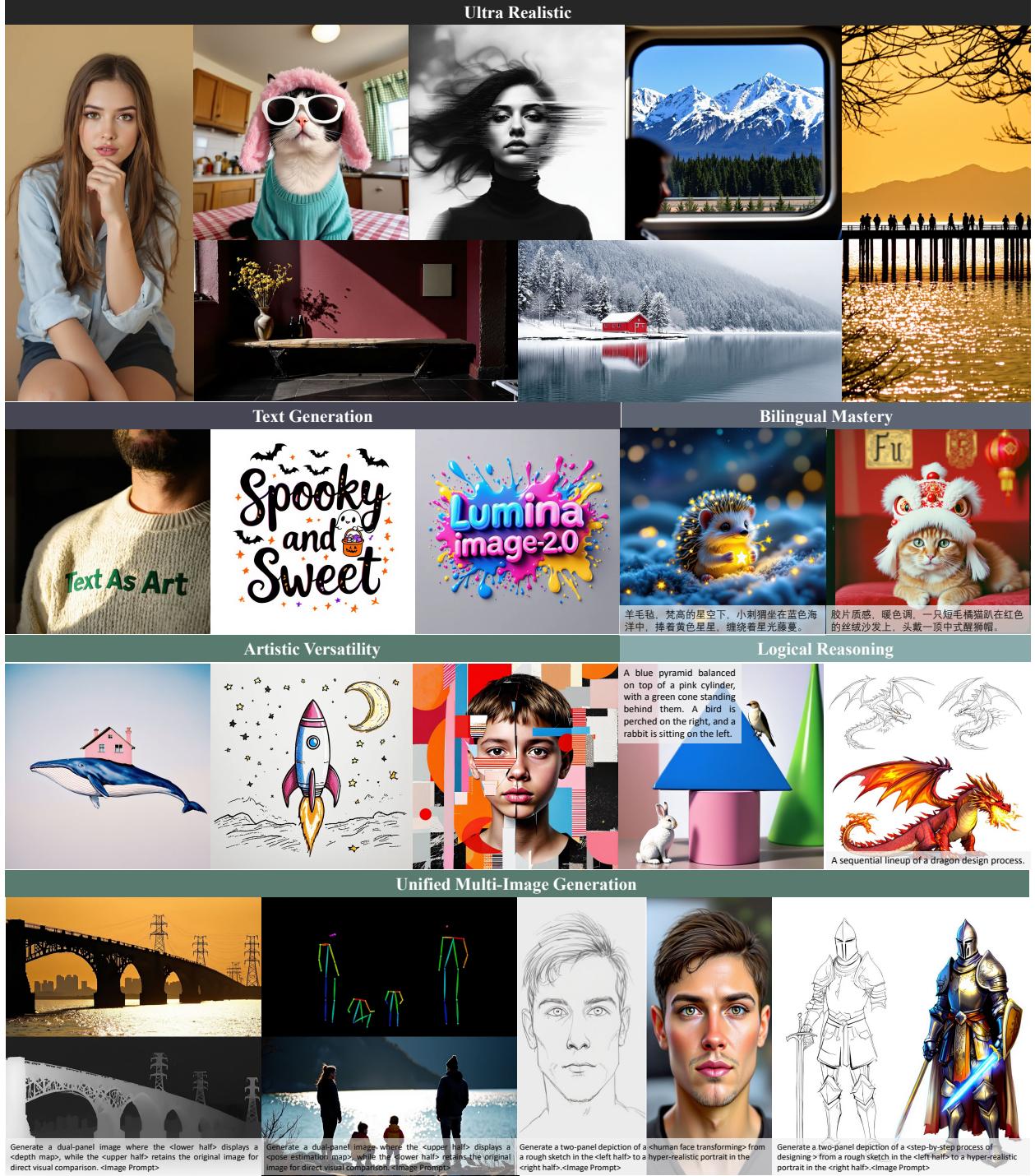


Figure 1: High-quality samples from our Lumina-Image 2.0, showcasing its capabilities in ultra-realistic, text generation, artistic versatility, bilingual mastery, logical reasoning, and unified multi-image generation.

Driven by the aforementioned challenges, we present **Lumina-Image 2.0**, a unified and efficient T2I generative framework that comprises four key components: (1) a Unified Next-DiT model for generating images that faithfully aligned with the text input, (2) a Unified Captioner (UniCap) for producing high-quality text-image pairs, and a series of specific designs for (3) efficient training and (4) efficient inference. Specifically, to address architectural limitations, our **Unified Next-DiT** model utilizes a joint self-attention mechanism, enabling our model to process both textual and visual tokens in a fully end-to-end manner, similar to decoder-only transformers in recent large

language models [23, 24, 25, 26]. This design facilitates seamless multimodal interaction, allowing for the integration of additional multimodal tokens or specific prompt templates to extend the model’s capabilities without modifying the core architecture. In response to the scarcity of high-quality textual descriptions in paired text-image data, we introduce **Unified Captioner (UniCap)**, a unified captioning system specifically designed for T2I generation. UniCap excels at precisely understanding complex scenes, and generating comprehensive and coherent multilingual descriptions. Leveraging these capabilities, we employ UniCap to create multi-granularity, multi-dimensional textual descriptions that better align with the images. Furthermore, our experiments reveal that when combining the unified Next-DiT and UniCap for training, the text-to-image attention in transformer blocks dynamically adjusts its capacity based on the length of textual embeddings, behaving similarly to a dynamic feed-forward network. This observation motivates us to further enhance model capacity and performance by increasing the richness of textual descriptions without introducing additional parameters.

Furthermore, both training and inference efficiency are crucial for model development and deployment. To perform **efficient training**, Lumina-Image 2.0 employs a multi-stage progressive training strategy with hierarchical high-quality data. The multi-domain system prompts and an auxiliary loss are further utilized to learn domain-specific knowledge and preserve low-frequency features, respectively. For **efficient inference**, we adopt several advanced sampling techniques and verify that the integration of CFG-Renormalization (CFG-Renorm) [27] and CFG-Truncation (CFG-Trunc) [28] can boost the sampling speed and maintain high sampling quality. Specifically, CFG-Renorm addresses the issue of over saturation at large classifier-free guidance (CFG) scales, while CFG-Trunc streamlines the inference process by eliminating redundant CFG computations, thereby enhancing both inference stability and speed. Additionally, we incorporate Flow-DPM-Solver [8] and TeaCache [29] to further optimize inference speed.

We evaluate Lumina-Image 2.0 on publicly available benchmarks, including DPG [30], GenEval [31], and T2I-CompBench [32]. Considering the limitation of current academic benchmarks in comprehensively evaluating T2I models, we report the ELO rankings of Lumina-Image 2.0 on several online T2I arenas, which is evaluated by human annotators. Our experimental results consistently demonstrate that Lumina-Image 2.0 achieves significant improvements over previous model (Lumina-Next [17]). We release the complete training details, code, and models to facilitate the full reproduction of Lumina-Image 2.0.

2 Related Work

Recent advancements in text-to-image generation have been remarkable. Diffusion-based models have progressively transitioned from U-Net architectures [33] to Diffusion Transformers [34], as demonstrated by models such as PixArt [15, 35], FLUX [9], SD3 [36], Lumina [16, 17], and SANA [8]. These Diffusion Transformers exhibit superior scalability and are progressively evolving toward a unified multimodal representation [37]. Regarding text encoders, early approaches [1] employed CLIP [38], while subsequent works [22, 36, 9] additionally adopted T5-XXL [39]. More recently, SANA [8], Lumina [17, 16] and our Lumina-Image 2.0 have incorporated Gemma [40] as the text encoder. Furthermore, the latest models leverage flow-based parameterizations [41, 42], which enhance both training and inference efficiency compared to conventional diffusion methods. In parallel, a range of advanced autoregressive and hybrid text-to-image models have emerged [43, 3, 6, 44, 10, 7], achieving performance on par with their diffusion-based counterparts. However, the sampling speed of these autoregressive models remains significantly slower than that of diffusion-based approaches, posing a critical challenge for their practical deployment.

Meanwhile, the advancement of text-to-image models has been significantly shaped by the evolution of vision-language models (VLMs) [45, 46, 47, 48], where the quality of image captions plays a critical role in both model performance [2, 36]. Currently, the most commonly employed image captioners in text-to-image research include LLaVA [49], CogVLM [50], ShareGPT-4 [46], and Qwen-VL [51, 52, 53], all of which are general-purpose vision-language models (VLMs). However, there is a significant lack of research focused on developing captioner models specifically tailored for the text-to-image task, which may impede the further advancement of text-to-image models.

3 Revisiting Lumina-Next

Model Architecture. Lumina-Next [17] introduces Next-DiT, a scalable flow-based diffusion transformer, as its core architecture. Building upon the original diffusion transformer [34], Next-DiT employs sandwich normalization and query-key normalization [54] to enhance training stability, and leverages 2D Rotary Positional Encoding [55] to encode positional information of images. For text-to-image generation, Next-DiT utilizes zero-initialized gated cross-attention to inject text embeddings extracted by Gemma [40].

Training and Inference Strategy. Lumina-Next is trained on approximately 20M synthetic text-image pairs, with image captions generated using user prompts and VLM models. During training, Lumina-Next employs a multi-stage

progressive training approach, similar to recent text-to-image models [15, 16]. This strategy involves sequential training at 256, 512, and 1024 resolutions, enabling the model to progressively capture both low-frequency and high-frequency information from images. During sampling, Lumina-Next introduces two time schedules tailored for flow models to minimize the ODE truncation errors, and it supports both first-order and higher-order solvers, such as Euler and Midpoint solvers.

Naive Data Scaling. Inspired by the data scaling paradigm of large-scale models [56, 24, 57, 36], we believe that the performance gap in Lumina-Next primarily stems from insufficient training data. Therefore, we scaled Lumina-Next’s dataset from 20M to 200M samples. This expanded dataset encompasses diverse real and synthetic data processed by the same cleaning and annotation pipeline. We retain the same model architecture and training strategy in Lumina-Next. We observed that the model performance shows considerable improvement across various academic metrics compared to Lumina-Next. For example, on the DPG benchmark [30], the performance improved from 75.66 to 85.80 after data scaling. This demonstrates the effectiveness of Next-DiT as a robust framework for scalable image generation.

4 Lumina-Image 2.0

4.1 Framework Overview

Lumina-Image 2.0 establishes a unified and efficient framework by integrating Unified Next-DiT, Unified Captioner (UniCap), and a set of efficient training and inference strategies. The overall pipeline is illustrated in Fig. 2, and we apply a custom filtering pipeline [15, 58] to select high-quality training images. To improve text quality, our UniCap re-captions the training data to generate accurate and detailed textual descriptions at multiple levels of granularity. The resulting high-quality image-text pairs are organized into a hierarchical training dataset, which is subsequently used to optimize the Unified Next-DiT model using our proposed training strategies. Finally, several inference strategies are further introduced to efficiently generate high-quality images.

4.2 Unified Next-DiT

After revisiting the architecture of Next-DiT, we observe that zero-initialized gated cross-attention for integrating text embedding limits the capability of text-image alignment and also requires additional architecture modification when adapting to new tasks. Therefore, we propose Unified Next-DiT, a unified text-to-image model that treats text and images as a unified sequence to perform joint attention inspired by recent advances in unified multimodal learning [37].

Architecture of Unified Next-DiT. Next-DiT employs Gemma [40] as the text encoder, whose text embeddings exhibit unidirectional positional bias [20] caused by the causal self-attention in the large language model. During generation, the biased text embeddings are fixed and sparsely injected to the transformer block via zero-initialized gated cross-attention. Therefore, we remove all zero-initialized gated cross-attention in Next-DiT. Instead, we leverage a unified single-stream block that fuses caption embeddings and noised latent by concatenating them and performing joint self-attention, which facilitates more effective text–image interaction and task expansion. As illustrated in Fig. 2, our single-stream blocks build upon the original DiT block with the addition of sandwich normalization and query-key normalization to ensure stable training. The Multimodal-RoPE [59] (mRoPE) is employed to jointly model text–image sequences in a unified manner, which encodes the text length as well as the image’s height and width into three dimensions. Moreover, we further observe that textual and visual features at the input level exhibit a considerable gap. To address this issue, we introduce text and image processors prior to the single-stream blocks. These processors with similar but lightweight single-stream blocks facilitate intra-modal information exchange and mitigate the gap between modalities. Since caption embeddings are fixed for all timesteps, the text processor does not incorporate timestep conditioning.

Comparison with previous architectures. As illustrated in Fig. 3, we compare the architecture of Unified Next-DiT with mainstream Diffusion Transformers. PixArt [15] and Lumina-Next [17] employ an additional cross-attention block after self-attention to inject fixed text embeddings, whereas our model adopts a single, unified attention module that jointly handles both text and noisy latent. Compared with the MMDiT architecture used in SD3 [36] and FLUX [9], the key difference is that MMDiT employs the double-stream blocks, allocating extensive and separate parameters for text and image sequences. In contrast, our method is designed from a more unified perspective, utilizing a single set of parameters to simultaneously model both the text and image sequences. Our model shares similarities with OmniGen [37], which introduces a single-stream causal DiT architecture for unified image generation. In pursuit of unifying the transformer architecture with auto-regressive models, OmniGen removes adaptive Layer Normalization (adaLN) and applies causal self-attention initialized from a large language model. However, adaLN is considered essential for Diffusion Transformers [34], and initializing from a language model may introduce conflicts with the knowledge for image generation.

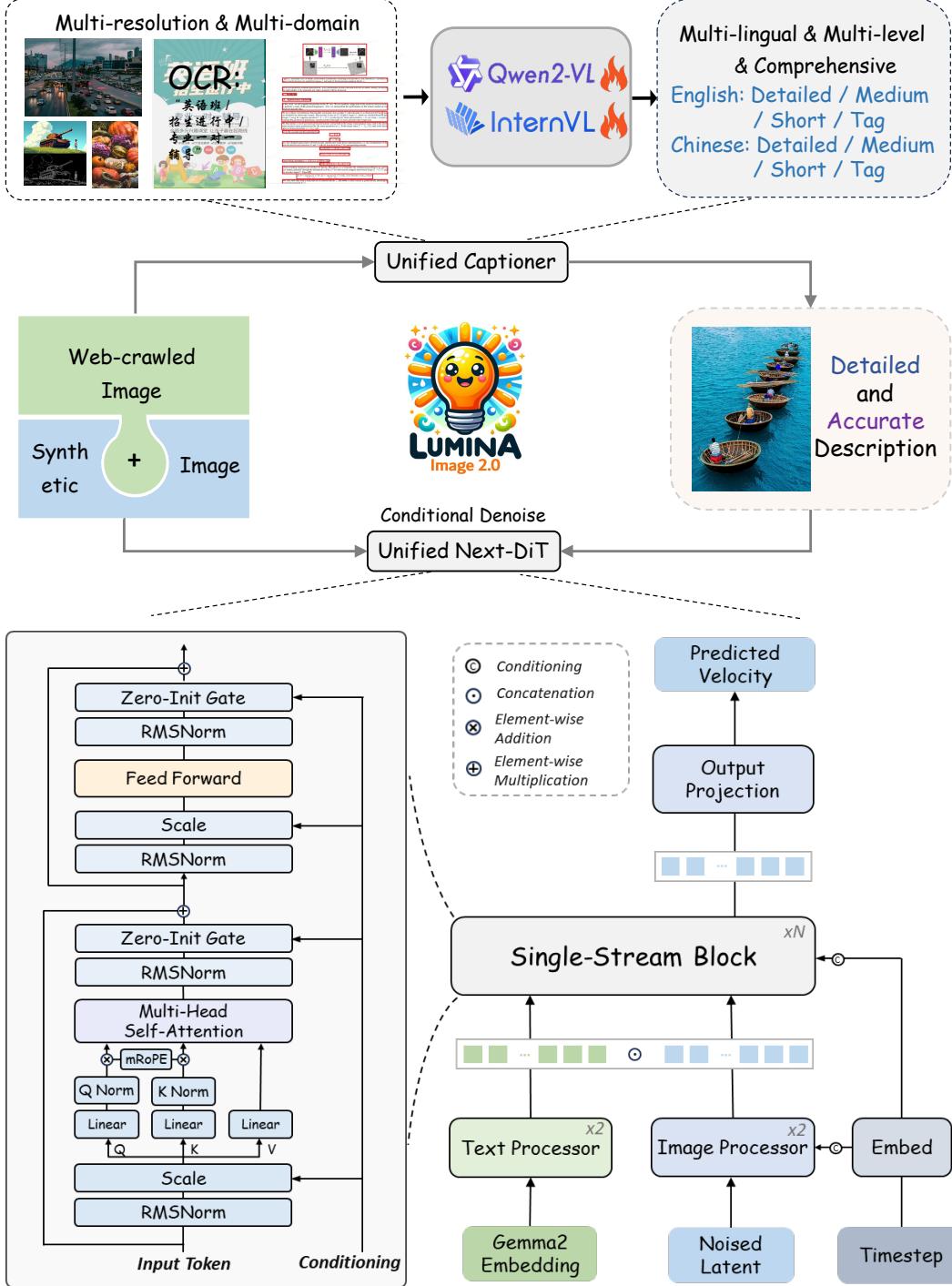


Figure 2: Overview of Lumina-Image 2.0, which consists of Unified Captioner and Unified Next-DiT. The Unified Captioner re-captions web-crawled and synthetic images to construct hierarchical text-image pairs, which are then used to optimize Unified Next-DiT with our efficient training strategy.

4.3 Unified Captioner

Due to the crucial role of image captions in enhancing model performance [2], using out-of-box pre-trained Vision Language Models (VLMs) for image recaptioning has been standard practice in previous literatures [22, 36, 15, 35,

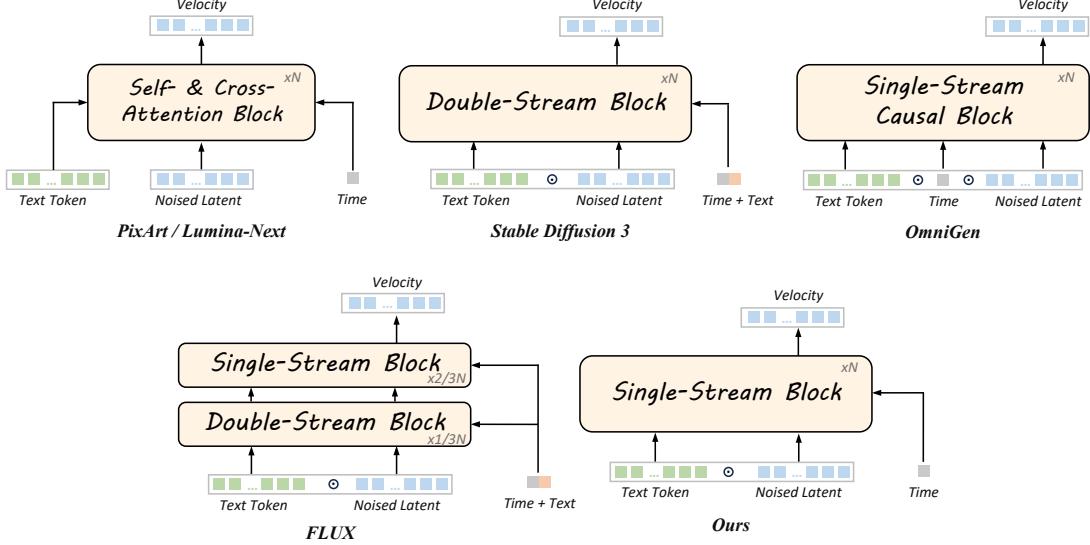


Figure 3: We compare the Diffusion Transformer architectures between our Unified Next-DiT, and PixArt [15], Lumina-Next [17], Stable Diffusion 3 [36], OmniGen [37] and FLUX [9].

16, 17]. However, these VLMs exhibit several limitations, including single-granularity descriptions, domain biases, and fixed low-resolution inputs, which result in suboptimal caption quality and a noticeable gap from real-world user prompts. To address these limitations and construct high-quality text-image datasets, we develop Unified Captioner (UniCap), a captioning system that unifies diverse visual inputs and provides multi-granularity, multi-perspective, and multi-lingual high-quality textual descriptions. In addition, we also introduce a unified perspective to make the caption-driven model capacity more interpretable.

Unifying Textual Description. To enable Lumina-Image 2.0 to handle diverse prompts – ranging from multi-granularity, multi-perspective, and multilingual descriptions – we train UniCap to deliver all types of descriptions to achieve unified image recaptioning. In particular, our approach comprises three key components: (1) For multi-granularity descriptions, we begin by carefully prompting GPT-4o [56] to generate highly detailed descriptions. Then we employ open-source large language models (LLMs) to simultaneously summarize detailed captions into medium, short, and tag-based descriptions for captioner training, which enables UniCap to deliver captions of multiple granularities while retaining essential information, as shown in Fig. 7 and Fig. 8. (2) For multi-perspective descriptions, we include image style descriptions, main object descriptions, all-object descriptions, object attribute descriptions, and spatial relationship descriptions, ensuring comprehensive coverage of visual elements, attributes, spatial structures, and stylistic nuances. (3) For multi-lingual descriptions, we utilize bilingual large language models to translate captions into Chinese, enabling UniCap to generate bilingual captions simultaneously. Surprisingly, although UniCap only captures all data in English and Chinese for Lumina-Image 2.0 training, the model benefits from Gemma’s multilingual capabilities and emerges with the understanding of other languages, thereby expanding its accessibility to a broader user base (see Fig. 6).

Unifying Visual Understanding. Existing VLMs struggle with processing images from diverse domains and open-world scenarios, and are limited to low-resolution inputs, making it difficult to capture fine-grained details of images. To address this issue, we train UniCap with a caption dataset that encompasses a wide range of visual content, including natural images, web-crawled images, photographs, synthetic images, multi-image documents, infographics, OCR-related images, and multilingual content, ensuring comprehensive domain coverage and conceptual diversity. Besides, unlike LLaVA [45] and ShareGPT4V [46], which resize images of varying scales and aspect ratios to a fixed low-resolution format, our UniCap processes images at their native scale in a unified manner. This approach yields more accurate and detailed captions, significantly reduces hallucinations, and improves OCR recognition. This strategy has been widely adopted by recent VLMs, including SPHINX-X [48], InternVL [47], and XComposer [60].

Furthermore, inspired by the concept of specialized generalist intelligence (SGI) [61], where AI systems excel in specialized tasks while maintaining broad general abilities, we aim for Lumina-Image 2.0 to not only showcase powerful text-conditioned generation capabilities but also serve as a unified interface for diverse visual generation tasks. To this end, we collect annotations from various visual tasks, such as depth maps, pose maps, canny maps, and sketches. We then concatenate them with paired images to form composite grids and leverage template captions (please refer to the last row of Fig. 1) to effectively describe the underlying logical process. This unified approach enables Lumina-Image

2.0 to handle advanced tasks beyond text-to-image generation, thereby laying the foundation for potential downstream applications.

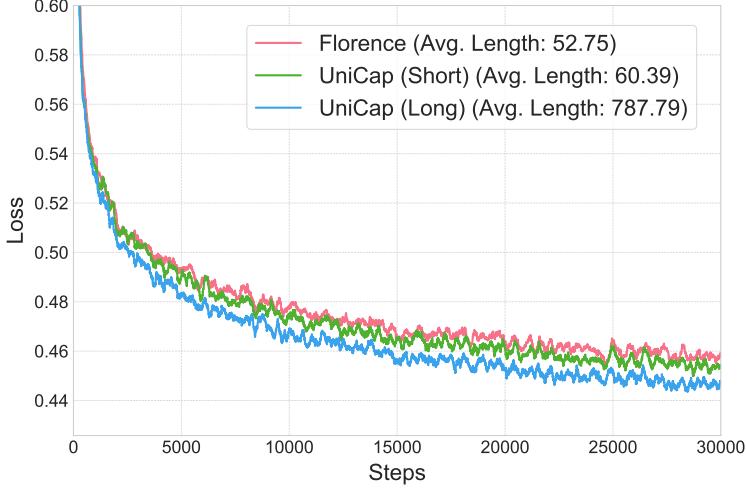


Figure 4: The training loss curve with respect to captions with different lengths. The “Avg. Length” represents the average character number.

A Unified Perspective on Caption-Driven Model Capacity. The importance of detailed image captions has been witnessed for scaling up diffusion models [2, 36]. During the training of Lumina-Image 2.0, we specifically observed that both the length and quality of image captions direct influence the model’s convergence speed. As shown in Fig. 4, we train the model using three versions of image captions: (1) short captions generated by Florence [62], (2) short but precise captions generated by our UniCap, and (3) long and detailed captions from UniCap. We observe that as captions become more precise and detailed, the model’s convergence speed significantly improved. During the inference phase, it is also commonly recognized by previous works [58, 22] that longer captions often lead to better generation results. These observations motivate us to rethink the role of caption embeddings in text-to-image generation. To further analyze the impact of image caption, in this paper, we provide an interpretable perspective on this phenomenon – the attention operation between texts and images can be viewed as a dynamic feed-forward network (FFN), where the choice of caption embeddings governs its effective knowledge integration and representational capacity.

Generally, the FFN layer of a transformer can be interpreted as a key-value memory that encapsulates the general knowledge acquired by the model [63], which can even be manually constructed without the need for training [64]. It also has been shown that the FFN layer can be effectively substituted by self-attention with persistent memory [65]. Motivated by these findings, we further explore the relationship between the text-to-image attention and the FFN mechanism. Note that the term “text-to-image attention” encompasses both the independent cross-attention used in Next-DiT [17] and PixArt [35], as well as the image-text interaction component in models performing joint self-attention (e.g., our Unified Next-DiT).

Given a sequence of image tokens $X \in \mathbb{R}^{L_{\text{img}} \times d}$ and a sequence of text tokens $Y \in \mathbb{R}^{L_{\text{text}} \times d}$. An ordinary image-text attention can be equivalently rewritten in the form of FFN as follows:

$$\text{Attn}(X, Y) = \sigma(X W_1(Y)) W_2(Y), \quad (1)$$

where $\sigma(\cdot)$ denotes the Softmax function, and the two “weight matrices” are conditioned on the text embeddings Y :

$$W_1(Y) = \frac{W_Q (Y W_K)^T}{\sqrt{d_k}} \in \mathbb{R}^{d \times L_{\text{text}}}, \quad (2)$$

$$W_2(Y) = Y W_V \in \mathbb{R}^{L_{\text{text}} \times d}, \quad (3)$$

where W_Q , W_K , and W_V are weight matrices for query, key, and value, respectively, and d_k is the dimension of query / key. Notably, the hidden dimension between $W_1(Y)$ and $W_2(Y)$ changes dynamically with the context length as L_{text} . Under this formulation, the text-to-image attention computation can be viewed as an FFN whose parameters are generated by a hyper-network, with *dynamic weights* and *dynamic hidden size*. Specifically, the conditional information (*i.e.*, the text) is encoded to form the dynamic weights, while the hidden size L_{text} will adjust the capacity of this FFN-like module via its length.

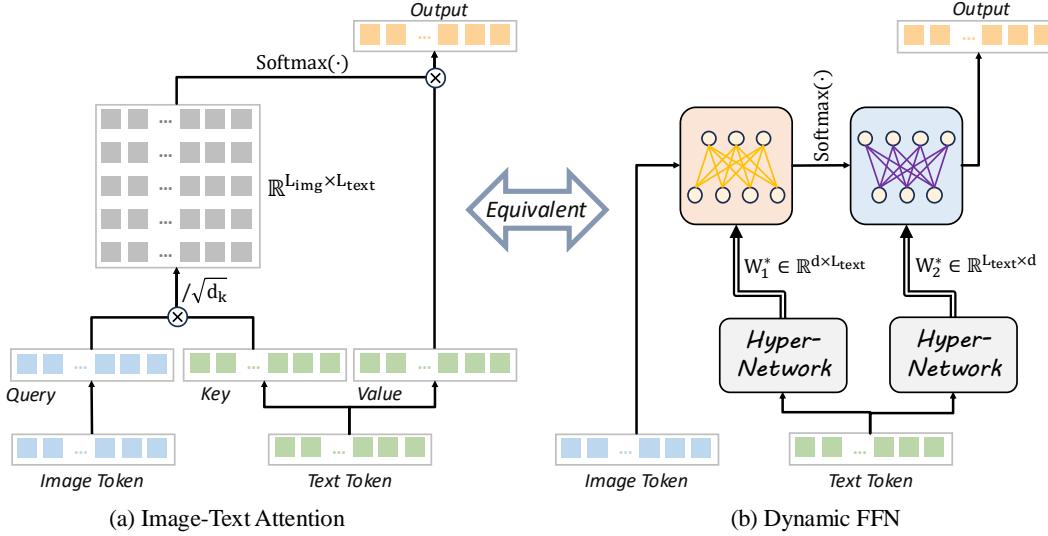


Figure 5: Illustration of reformulating the Image-Text Attention as an FFN generated by a hyper-network, with its weights and hidden dimensions dynamically determined by the input text token.

From this perspective, we reach an interesting conclusion – increasing caption length effectively serves as a controllable means of scaling up model parameters. This insight suggests that the capacity of the model in both training and inference can be modulated simply by adjusting the length of the caption, which can lead to improved knowledge learning and overall performance. These findings are consistent with recent trends in existing work [66, 67] and highlight promising directions such as inference-time scaling [68].

4.4 Efficient Training

Lumina-Image 2.0 introduces an efficient training framework that integrates multi-stage progressive training, a hierarchical high-quality dataset, multi-domain system prompts, and auxiliary loss. These strategies improve image quality and detail refinement while accelerating convergence.

Multi-Stage Progressive Training. Unlike prior approaches [15, 33, 17, 37] that optimize generative models over three progressive resolution stages, we skip the intermediate 512 resolution stage and introduce an additional high-quality tuning phase. This results in a three-stage progressive training pipeline: a low-resolution phase (256 resolution), a high-resolution phase (1024 resolution), and a high-quality tuning phase (1024 resolution). In the low-resolution phase, Lumina-Image 2.0 focuses on learning global and low-frequency information, such as domain knowledge, object relationships, and structural patterns. The subsequent two phases first transfer this knowledge to higher resolutions and further enhance fine-grained visual details.

Hierarchical High-quality Data. In contrast to Lumina-Next, which utilizes a fixed dataset in all training stages, we construct a hierarchical dataset by filtering images based on image quality criteria (e.g., aesthetic) at different stages. Specifically, we begin with a dataset of 110M samples. For the low-resolution training stage, we select 100M samples. The remaining 10M samples, containing relatively higher-quality data, are then used in the high-resolution phase. From these, we further curate a subset of the highest-quality 1M samples for the final fine-tuning stage.

Multi-domain System Prompt. We collect training data from diverse domains, including high-aesthetic synthetic data, as well as photorealistic real-world data. However, there is a substantial domain gap between these datasets, often resulting in slower convergence and difficulties in learning domain-specific knowledge. Motivated by ChatGPT [23], we propose distinct system prompts to differentiate between these domains, thereby reducing learning difficulty and accelerating convergence. Specifically, during our proposed three-stage progressive training phase, we use two types of system prompts (“Template A” and “Template B”) that are directly prepended to the image prompt, as shown in Tab. 1. For the unified multi-image generation, we introduce an additional fine-tuning phase (see Sec. 5.1 for details) with the system prompt “Template C”.

Auxiliary Loss. When training our model on high-resolution images, the model exhibits significant improvements in high-frequency details while some degradations in low-frequency structures. We introduce an auxiliary loss to address

Table 1: Prompt template for Lumina-Image 2.0. **<Image Prompt>** will be replaced with the user specific image description. **<lower half>** and **<upper half>** will be replaced with the specific spatial relationships. **<depth map>** will be replaced with the target image type.

Template A	You are an assistant designed to generate high-quality images based on user prompts. <Prompt Start> <Image Prompt>
Template B	You are an assistant designed to generate superior images with the superior degree of image-text alignment based on textual prompts or user prompts. <Prompt Start> <Image Prompt>
Template C	Generate a dual-panel image where the <lower half> displays a <depth map> , while the <upper half> retains the original image for direct visual comparison. <Prompt Start> <Image Prompt>

this issue, which computes the flow-matching objective [42] with latent features downsampled by a factor of 4:

$$\mathcal{L}_{\text{aux}}(\theta) = \mathbb{E}_{t,x,\epsilon} \|v_\theta(x_t, t) - u_t\|^2, \quad (4)$$

where $t \in [0, 1]$ denotes timestep, $x = \text{AvgPool}_4(z)$ denotes the downsampled latent features using average pooling by a factor of 4, $\epsilon \sim \mathcal{N}(0, I)$ is random Gaussian noise, $v_\theta(\cdot)$ and $u_t = x - \epsilon$ represent the predicted vector field and target vector field, respectively. This approach helps preserve low-frequency features while learning high-frequency details, enabling efficient knowledge integration and allowing direct fine-tuning at 1024 resolution.

4.5 Efficient Inference

To boost the sampling speed as much as possible while maintaining high sampling quality, Lumina-Image 2.0 makes a deeper exploration on inference efficiency.

CFG-Renormalization (CFG-Renorm). Classifier-free guidance (CFG) [69] is known for improving both visual quality and text-image alignment. During inference, at each timestep t , the predicted velocity v_t is calculated as $v_t = v_{t_u} + w(v_{t_c} - v_{t_u})$, where w is the CFG scale, v_{t_c} and v_{t_u} represent the conditional and unconditional velocity, respectively. However, scaling by a large w may introduce extremely high activations in certain dimensions of v_t , and these abnormal values can result in visual artifacts in the final generated samples. To address this, recent work introduces the CFG-Renorm method [27] to rescale the magnitude of the modified velocity of v_t using that of the conditional velocity v_{t_c} . We find that this technique effectively improves the stability of CFG-guided generation without introducing additional computational costs.

CFG-Truncation (CFG-Trunc). Recent research [28] indicates that text information is largely captured in the early generation stages. Therefore, evaluating v_{t_c} beyond the early timesteps may be redundant. The CFG-Trunc can be formulated as follows:

$$v_t = \begin{cases} v_{t_u} + w(v_{t_c} - v_{t_u}) & t \geq \alpha; \\ v_{t_u} & t < \alpha. \end{cases} \quad (5)$$

where α denotes a predefined threshold. This modification can achieve over a 20% acceleration in sampling speed, without visual degradation.

Flow-DPM-Solver (FDPM). Lumina-Next supports a range of ODE solvers, such as Midpoint and Euler method. While these solvers ensure stability, they are relatively slow since they are not designed for flow models, requiring a large number of function evaluations (NFE) for convergence. To improve this, we integrate FDPM [8, 70], which modify DPM-Solver++ [70] to flow models, into Lumina-Image 2.0. FDPM achieves convergence in just 14-20 NFEs, providing a faster sampling solution. However, we find that FDPM sometimes suffers from poor stability in practice.

Timestep Embedding Aware Cache (TeaCache). TeaCache [29] is designed to selectively cache informative intermediate results during the inference, thereby accelerating diffusion models. TeaCache has successfully accelerated various mainstream image and video generation models, including FLUX [9], HunyuanVideo [58], as well as Lumina-Next. Building on its success, we integrate TeaCache into Lumina-Image 2.0. However, our experiments show that TeaCache also introduces visual quality degradations when combined with the above techniques.

Discussion. The above four inference strategies are mutually compatible and can be applied in combination. Notably, Lumina-Image 2.0 is the first to demonstrate that CFG-Renorm and CFG-Trunc provide complementary benefits when

applied together. CFG-Renorm addresses the issue of over-saturation and visual artifacts when the CFG scale is large, while CFG-Trunc further alleviates this problem by eliminating redundant CFG calculations and achieving acceleration at the same time. The flexibility of the CFG scale can be significantly extended to a wider range by combining these techniques. FDPM and TeaCache can also be integrated into the pipeline, but both of them present certain challenges. FDPM lacks sufficient stability and frequently produces suboptimal samples while TeaCache results in blurriness in the sampled images. For further details, refer to Sec. 5.4.

5 Experiments

5.1 Implement Details

Training Dataset. Following the methods in [3, 4, 7, 36, 15, 37, 71], we constructed a dataset combining both real and synthetic data, and performed data filtering based on the techniques outlined in [15, 22, 58], resulting in total 110M samples. This dataset is reorganized into three training phases, with 100M, 10M, and 1M samples for each training phase. As the dataset size decreased, the quality of the data progressively improved.

Architecture and Training Setups. The architecture configurations of our Unified Next-DiT model, along with a comparison to Lumina-Next [17], are summarized in Tab. 2. We employed 32 A100 GPUs across all three stages to optimize our Unified Next-DiT. The corresponding training configurations are detailed in Tab. 3. In addition, for multi-image generation task, we introduce an extra fine-tuning phase, where we consolidate different visual tasks into image grids and generate captions for these concatenated grids to form image-pair pairs. Besides, for the UniCap model, we finetune the Qwen2-VL-7B [59] based on the constructed caption dataset with multi-domain visual data and diverse textual descriptions.

Table 2: Comparison of configuration between Lumina-Next and Lumina-Image 2.0.

Model	Params	Patch Size	Dimension	Heads	KV Heads	Layers	RMSNorm ϵ [72]	Pos. Emb.
Lumina-Next	1.7B	2	2304	16	8	24	$1e^{-5}$	2D-RoPE
Lumina-Image 2.0	2.6B	2	2304	24	8	26	$1e^{-5}$	M-RoPE

Table 3: Training configuration across different stages.

Stage	Image Resolution	#Images	Training Steps (K)	Batch Size	Learning Rate	GPU Days (A100)	Optimizer
Low Res. Stage	256×256	100M	144	1024	2×10^{-4}	191	
High Res. Stage	1024×1024	10M	40	512	2×10^{-4}	176	AdamW [73]
HQ Tuning Stage	1024×1024	1M	15	512	2×10^{-4}	224	

5.2 Quantitative Performance

Main Results. We evaluate our model on three benchmarks: GenEval [31], DPG [30], and T2I-CompBench [32]. As shown in Tab. 4, Our model demonstrates strong performance across various metrics on the GenEval benchmark. In the Two Object, Counting, Color Attribute, and Overall metrics, we achieve the second-best performance compared to autoregressive and diffusion models. On the DPG benchmark, Lumina-Image 2.0 outperforms all compared models across three sub-metrics (Entity, Relation, and Attribute) as well as the Overall metric. Similarly, on T2I-CompBench, our model achieves the best results in both Color and Shape metrics. The significant advantage we achieve on the DPG benchmark is attributed to the detailed and accurate captions curated by our carefully designed captioning system. Our UniCap generates extremely long and detailed descriptions, which align with the characteristics of prompts contained in DPG, resulting in the strong performance across various metrics, especially in the Relation score.

Comparison with ELO Scores. To better evaluate our model, we present evaluation results from three text-to-image arenas, with all ELO scores [76] based on ratings from human annotators. (1) We first perform tests on Artificial Analysis⁴. As shown in Tab. 5, Lumina-Image 2.0 achieves mid-tier results, outperforming almost all open-source models (e.g., SD3 [36] and Janus Pro [10]) and several closed-source systems (e.g., DALL-E 3 [2]), but still lags behind the top closed-source models, such as FLUX Pro [9]. (2) To analyze the alignment and coherence abilities of our model, we also provide rankings from Rapidata⁵. As shown in Tab. 6, our model achieves a comparable ranking.

⁴<https://artificialanalysis.ai/text-to-image/arena?tab=Leaderboard>

⁵<https://www.rapidata.ai/leaderboard/image-models>

Table 4: Performance comparison across different models on GenEval [31], DPG [30], and T2I-CompBench [32] benchmarks. "↓" or "↑" indicate lower or higher values are better. **Bold** indicates the best performance, while underlining denotes the second-best performance.

Methods	# Params	GenEval ↑				DPG ↑				T2I-CompBench ↑		
		Two Obj.	Counting	Color Attri.	Overall	Entity	Relation	Attribute	Overall	Color	Shape	Texture
AutoRegressive Models												
LlamaGen [4]	0.8B	0.34	0.21	0.04	0.32	-	-	-	65.16	-	-	-
Chameleon [5]	7B	-	-	-	0.39	-	-	-	-	-	-	-
HART [6]	732M	-	-	-	-	-	-	-	80.89	-	-	-
Show-o [3]	1.3B	0.52	0.49	0.28	0.53	-	-	-	67.48	-	-	-
Emu3 [7]	8.0B	0.81	0.49	0.45	0.66	87.17	90.61	86.33	81.60	0.7544	0.5706	0.7164
Infinity [44]	2B	0.85	-	0.57	0.73	-	90.76	-	83.46	-	-	-
Janus-Pro-1B [10]	1.5B	0.82	0.51	0.56	0.73	88.63	88.98	88.17	82.63	-	-	-
Janus-Pro-7B [10]	7B	0.89	0.59	0.66	0.80	88.90	89.32	<u>89.40</u>	84.19	-	-	-
Diffusion Models												
LDM [1]	1.4B	0.29	0.23	0.05	0.37	-	-	-	63.18	-	-	-
SDv1.5 [1]	0.9B	-	-	-	0.40	74.23	73.49	75.39	63.18	0.3730	0.3646	0.4219
Lumina-Next [17]	1.7B	0.49	0.38	0.15	0.46	83.78	89.78	82.67	75.66	0.5088	0.3386	0.4239
SDv2.1 [1]	0.9B	0.51	0.44	0.50	0.47	-	-	-	68.09	0.5694	0.4495	0.4982
PixArt- α [15]	0.6B	0.50	0.44	0.07	0.48	79.32	82.57	78.60	71.11	0.6886	0.5582	0.7044
SDXL [33]	2.6B	0.74	0.39	0.23	0.55	82.43	86.76	80.91	74.65	0.6369	0.5408	0.5637
SD3-medium [36]	2B	0.74	0.63	0.36	0.62	91.01	80.70	88.83	84.08	-	-	-
JanusFlow [74]	1.3B	0.59	0.45	0.42	0.63	87.31	89.79	87.39	80.09	-	-	-
Sana-0.6B [8]	0.6B	0.76	0.64	0.39	0.64	89.50	90.10	89.30	83.60	-	-	-
Sana-1.6B [8]	1.6B	0.77	0.62	0.47	0.66	<u>91.50</u>	<u>91.90</u>	88.90	<u>84.80</u>	-	-	-
DALL-E3 [2]	-	0.87	0.47	0.45	0.67	89.61	90.58	88.39	83.50	<u>0.8110</u>	0.6750	0.8070
OmniGen [37]	3.8B	0.86	0.64	0.55	0.70	-	-	-	-	-	-	-
Sana-1.5 [67]	4.8B	0.85	0.77	0.54	0.72	-	-	-	85.00	-	-	-
Lumina-Image 2.0	2.6B	<u>0.87</u>	<u>0.67</u>	<u>0.62</u>	<u>0.73</u>	91.97	94.85	90.20	87.20	0.8211	<u>0.6028</u>	<u>0.7417</u>

Table 5: Comparison of ELO scores evaluated in text-to-image arena from Artificial Analysis⁴ (as of February 23, 2025).

Methods	Artificial Analysis			
	Overall	Traditional Art	Fantasy & Mythical	Anime
FLUX1.1 [pro] [9]	1122	1075	1111	1127
FLUX1 [pro] [9]	1107	983	1081	1051
Lumina-Image 2.0	982	1015	1051	1037
DALLE 3 [2]	970	1008	1026	977
SD3 Medium [36]	945	990	1026	929
Janus Pro [10]	748	828	784	766

Table 6: Comparison of ELO scores evaluated in text-to-image arena from Rapidata⁵ (as of February 23, 2025).

Methods	Rapidata		
	Overall	Alignment	Coherence
FLUX1.1 [pro] [9]	1040	1036	1023
Imagen 3 [75]	1018	1003	1032
Lumina-Image 2.0	969	1031	986
DALLE 3 [2]	952	1022	958
SD3 Medium [36]	952	1022	984
Janus Pro [10]	734	932	947

In particular, Lumina-Image 2.0 ranks second only to FLUX Pro in terms of prompt alignment, exceeding many other closed-source models such as Imagen 3 [75]. This further validates the effectiveness of our proposed Unified Next-DiT architecture and UniCap annotation system. Moreover, although Janus Pro[10] achieves state-of-the-art results on academic benchmarks, its scores were considerably lower than those of Lumina-Image 2.0 and FLUX Pro on user-driven leaderboards. This discrepancy highlights the inherent bias and limitations in current academic benchmarks. (3) Finally, results from AGI-Eval⁶ in Tab. 7 demonstrate that Lumina-Image 2.0 significantly outperforms the previous Lumina-Next [17] as well as all other Chinese open-source models [77, 22].

In summary, we hope that this comprehensive evaluation and comparative analysis will provide the community with a clearer understanding of Lumina-Image 2.0's capabilities and constraints, thereby guiding future improvements. We also believe that developing better human-aligned evaluation benchmarks is essential to accurately assess current models and advance generative modeling progress.

⁶<https://ai-ceping.com/>

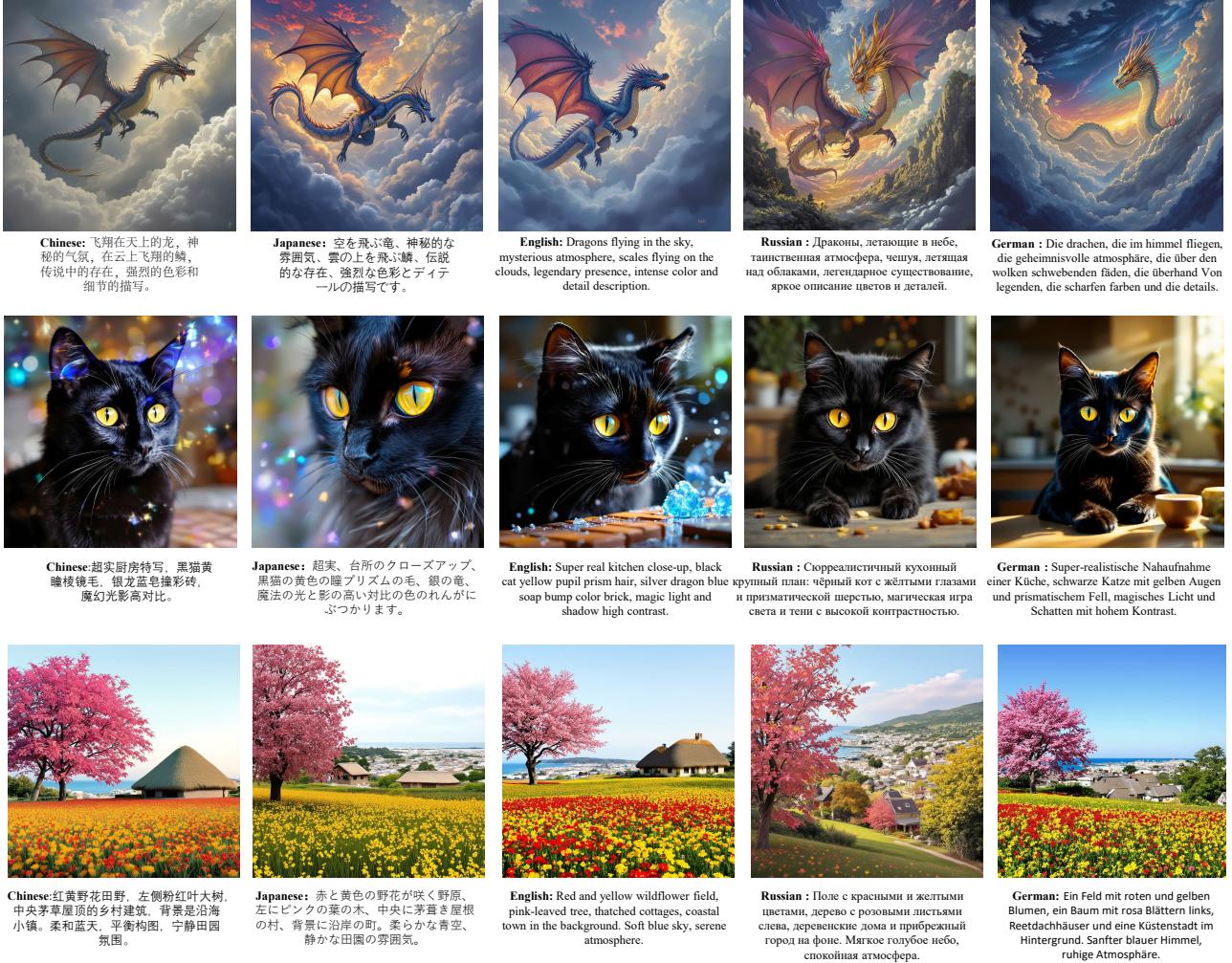


Figure 6: Visualization results of multilingual text-to-image generation by our Lumina-Image 2.0, covering five languages: Chinese, Japanese, English, Russian, and German.

Table 7: Comparison of ELO scores evaluated in text-to-image arena from AGI-Eval⁶ (as of February 23, 2025).

Model	FLUX1.1 [pro] [9]	FLUX.1 [dev] [9]	Lumina-Image 2.0	Kolors [77]	HunyuanDiT [22]	Lumina-Next [17]
Score	0.4859	0.4712	0.4545	0.3924	0.3920	0.3229

5.3 Qualitative Performance

Multi-lingual Generation. Compared to previous T2I models [15, 78] that use CLIP [38] and T5 [39] as text encoders, we employ Gemma2-2b [79] as the text encoder, enabling our model to understand multiple languages. It naturally exhibits zero-shot capability in languages such as German, Japanese, and Russian. As shown in Fig. 6, we present the generation results in five different languages.

Captioning Everything With UniCap. We compare our proposed UniCap with existing captioners, such as ShareGPT4V [46] and Florence [62], from four dimensions: complex scenes, dense text, visual understanding, and spatial relationships. UniCap supports multilingual annotations, including both Chinese and English, and can generate captions of varying lengths to accommodate diverse user needs. As shown in Fig. 7 and Fig. 8, UniCap delivers highly detailed and accurate descriptions, significantly outperforming the other two methods.

High-quality Image Generation. In Fig. 9, we present additional generation results of Lumina-Image 2.0. These results illustrate that our model is capable of producing high-quality images in various resolutions that are visually



UniCap (Ours)



ShareGPT4V



Florence



Detailed: The image captures a bustling scene in a museum, where the grandeur of history and the awe of artistry are on full display. The museum's high ceiling, adorned with ornate decorations, houses a large skylight that bathes the space in natural light. The walls, painted in pristine white, serve as a canvas for an array of paintings that add a splash of color to the otherwise monochrome setting. The floor, made of polished wood, reflects the light from above, creating a warm and inviting atmosphere. Hanging from the ceiling are chandeliers, their crystals catching the light and casting a soft glow throughout the room. The museum is filled with exhibits that draw the attention of the visitors. Dinosaur skeletons, their bones meticulously arranged, offer a glimpse into the prehistoric era. Paintings, each telling a story of its own, are arranged on the walls, inviting viewers to lose themselves in the world of art. People are scattered throughout the museum, some standing in quiet contemplation, others walking around, their curiosity piqued by the exhibits. Their actions bring life to the museum, making it more than just a collection of artifacts; it's a living, breathing space where history and art come to life.

Medium: Wide-angle photograph of an expansive, ornate museum hall with a high, arched ceiling featuring a grid of glass panels. The hall is filled with dinosaur skeletons displayed on wooden platforms, arranged in rows along the length of the room. The central focus is a large, dark-colored dinosaur skeleton prominently positioned in the middle of the hall. Visitors, including a woman in a blue shirt and a man in a green shirt, are scattered throughout the space, some standing on the upper balcony level, others on the ground level. The upper walls are adorned with framed artworks and decorative sculptures. The ceiling's intricate design and the warm ambient lighting create a grand and inviting atmosphere. The perspective is slightly elevated, capturing the grandeur and symmetry of the space. Photorealistic, high contrast, warm color palette, museum interior, detailed textures, balanced composition, historical and educational ambiance.

Short: The Museum of Natural History in Paris, France, boasts an impressive interior with a high ceiling adorned with a grid of lights and intricate architectural details. The walls are adorned with framed artworks and artifacts, while the floor is lined with wooden platforms displaying various dinosaur skeletons. Visitors can be seen walking around, admiring the exhibits. The floor is made of wooden planks, and the overall ambiance of the museum is one of historical significance and educational value.

Tag: museum interior, dinosaur skeletons, ornate architecture, high ceiling, natural light, educational, grand, symmetrical composition, vibrant colors

Detailed: 在法国巴黎的自然历史博物馆内，一个宏伟的室内空间焕发生机，展示着令人叹为观止的恐龙骨架陈列。这个宽敞的房间拥有高高的天花板，装饰着精致的拱门和天窗，让自然光线洒满整个空间。墙壁上装饰着各种各样的艺术品和信息展示，而地面则铺设着木制地板。恐龙骨架被精心地摆放在木制平台上，吸引着游客探索和欣赏。游客们四处漫步，有些人停下来仔细观察，而有些人则继续他们的探索之旅。尽管这幅画面非常美丽，但其实所提供的信息的准确性至关重要。

Medium: 展现在眼前，其天花板装饰着精致的拱门和玻璃面板，让自然光线得以过滤。墙壁上装饰着各种各样的艺术品和信息展示，而中央的走道则陈列着引人入胜的化石和雕塑。游客们沿着走道漫步，仔细观察展品，为这个空间增添了活力。尽管这幅图像展示了博物馆的美丽，但核实时所提供的信息的准确性至关重要。

Short: 在法国巴黎的自然历史博物馆内，一个宏伟的室内空间展现在眼前，其天花板装饰着精致的拱门和玻璃面板，让自然光线得以过滤。墙壁上装饰着各种各样的艺术品和信息展示，而中央的走道则陈列着引人入胜的化石和雕塑。游客们沿着走道漫步，欣赏着历史和科学的奇迹。

Tag: 博物馆内部，骨骼展览，高天花板，自然光，木制平台，精美装饰，恐龙骨架，人类骨架，游客，壁画，建筑摄影，广角镜头，对称构图，温暖色调，历史氛围

Short (Florence): The Museum of Natural History in Paris features a grand interior with a high ceiling, framed artworks, and wooden platforms showcasing dinosaur skeletons, where visitors explore and admire the exhibits.



Short (Florence): A promotional poster for online English learning classes features a young man with a pencil and book, bold text, and a call-to-action button on a dark blue modern background.

Detailed: Promotional poster for online English learning classes, featuring a young man with short brown hair, smiling and holding a pencil to his chin, positioned on the right side of the image. He wears a dark blue t-shirt and holds an open book in his left hand. The background is a dark blue with circular bokeh patterns. On the left side, there is a large speech bubble containing text in white and blue fonts. The text reads "Online English Learning Classes," "Certified English Language Teacher with over 5 years of experience," and "Register Now." Additional details include "Mon-Thu Day 0:00 AM-1:00 PM," a website URL "www.yourdomain.com," and a contact number "+000 0000 0000." A logo placeholder is present in the top right corner. The design is modern and clean, with a focus on clarity and readability, using a limited color palette of blue, white, and black, digital graphic design, high contrast, professional, educational, inviting.

Medium: Promotional poster for online English learning classes featuring a young man in a blue shirt holding a pencil and an open book, set against a dark blue background with circular patterns. The text highlights a certified teacher with over five years of experience, a "REGISTER NOW" call-to-action, and details about class schedule and contact information. The design is modern, professional, and emphasizes education.

Short: A promotional poster for online English learning classes features a young man in a blue shirt holding a pencil and an open book, with bold text and a call-to-action button, set against a dark blue background with a modern, professional design.

Tag: Online English learning, advertisement, young man, dark blue shirt, yellow pencil, open book, dark blue background, white speech bubble, text, registered teacher, experience, call to action, registration, contact number, logo, modern design, vibrant colors, professional, engaging, promotional material.

Detailed: 在线英语学习课程宣传图，图中是一位年轻男子，他留着短发，身穿蓝色T恤，右手拿着一本打开的书，左手拿着铅笔，下巴搁在铅笔上，目光向上，表情沉思。背景为深蓝色，带有圆形图案。左侧有一个白色对话框，里面包含文字：“在线英语学习课程，认证英语语言教师，拥有超过5年经验！立即注册，周一至周四，0:00 AM-1:00 PM，www.yourdomain.com，电话+000 0000 0000”。右上角有“LOGO HERE”字样。设计风格简洁现代，色彩鲜艳，对比强烈，字体清晰易读，专业且具有教育意义。

Medium: 在线英语学习课程的广告展示了一位穿着蓝色T恤的年轻男子，他手持铅笔和笔记本，背景是深蓝色的圆形图案。文字突出了“在线英语学习课程”和“超过5年经验的认证英语语言教师！”并提供了一个“立即注册”按钮。还包含了营业时间和网站链接，以及一个电话号码供咨询。设计简洁、现代且专业，突出了教育内容。

Short: 一个宣传在线英语学习课程的广告，展示了一位年轻男子手持铅笔和书籍，背景为深蓝色，文字为白色和蓝色，提供注册详情、教师资质和联系方式。

Tag: 在线英语课程，宣传海报，年轻男子，教育广告，现代设计，鲜艳色彩

Figure 7: Comparison with ShareGPT4V [46] and Florence [62] in complex scenes and dense text for caption generation. The **blue underline** correspond to areas with more detailed and accurate descriptions, while **red underline** and **red strikethrough** represent the incorrect and insufficient descriptions respectively.

realistic, highly aesthetic, and creatively expressive. Furthermore, extensive experiments with both Chinese and English prompts of different lengths demonstrate robust text-image alignment.

5.4 Ablation Study

Ablation Study on Multi-Stage Training Strategy. During our three-stage progressive training, the model’s performance steadily improves as dataset size decreased and quality increased. As shown in Tab. 8 and Fig. 10, the performances continue to advance from the second to the third stage, evidenced by both quantitative improvements

13


UniCap (Ours)



ShareGPT4V

Florence



Detailed: Top image: A close-up photograph of a plate with sliced figs arranged in a circular pattern on a wooden table. The figs are halved, revealing their vibrant pink and red interiors with a purple outer skin. A small bunch of fresh green leaves, possibly arugula, is placed in the center. The plate is a rustic blue ceramic with a brown rim. The lighting is natural, coming from the left, casting soft shadows and highlighting the textures of the figs and leaves. Photorealistic, shallow depth of field, warm color palette, high contrast, natural lighting, rustic and organic feel, visually balanced composition.

Medium: A top-down view of a blue ceramic plate with halved figs in a circular pattern, adorned with fresh green leaves, sits on a wooden table. The figs display vibrant pink insides and purple skin. Below, a detailed black and white sketch of figs and leaves on a black background provides a minimalist contrast. The scene features natural lighting, a warm color palette, and a balanced composition, highlighting the contrast between the photorealistic image and the sketch.

Short: A top-down view of a plate with halved figs and fresh greens on a wooden table, accompanied by a detailed black and white line drawing of figs and branches, showcasing a contrast between realistic photography and minimalist illustration.

Tag: Fig, blue ceramic plate, wooden table, green leaves, vibrant, natural lighting, shallow depth of field, rustic, artistic, minimalistic, monochrome sketch, botanical illustration, contrast, texture, organic, elegant, visually balanced

Detailed: 俯视图。一个陶瓷盘子装满了切片无花果，放在木桌上。盘子位于中心，上面有六个无花果，每个都展示了鲜艳的粉红色果肉和紫色外皮。新鲜的绿色叶子，可能是芝麻菜，被放在无花果中间。图片的下半部分是一个黑白的无花果切片和叶子的线条画，背景是黑色的。这幅画捕捉了无花果的复杂细节和叶子的纹理。自然光线，浅景深，温暖的色彩调色板，摄影写实，高对比度，乡村风格，有机感觉，艺术构图，视觉平衡。

Medium: 这张图片分为两部分。上半部分展示了一个充满活力的静物画，画中切开的无花果放在一个蓝色的陶瓷盘子上，摆放在一张木制桌子上，中间点缀着新鲜的绿叶；下半部分则是一个黑白的无花果切片素描，背景为黑色，突出了无花果的纹理和形状。整个构图将现实主义与插画相结合，通过对比鲜明的色彩和元素，创造出一个视觉上引人入胜且平衡的画面。

Short: 一个充满活力的无花果切片在蓝色陶瓷盘子上的俯视图，配有新鲜的绿叶，放置在木制桌子上，下方有黑白素描，营造出一种视觉上平衡且对比强烈的构图。

Tag: 无花果，食物摄影，木桌，自然光，高对比度，乡村风格，极简主义，艺术对比



Short (Florence): A top-down view of a plate with halved figs and greens on a wooden table, alongside a black and white line drawing, highlighting the contrast between realism and minimalism.

Detailed: The image presents a striking contrast between two different perspectives of a plate of figs. In the top photo, a blue plate cradles ~~six figs~~, their vibrant red and purple hues standing out against the cool blue of the plate. The figs are arranged in a circular pattern, creating a sense of symmetry and balance. ~~A sprig of parsley~~ rests atop the figs, adding a touch of green to the otherwise ~~monochromatic palette~~. The bottom photo is a black and white sketch of the same plate of figs. The absence of color focuses the viewer's attention on the shapes and forms within the composition. The figs, still life, and parsley are sketched with varying degrees of detail, creating a layered effect that adds depth to the image. ~~Despite the absence of color, the image is rich in detail and contrast, offering a unique perspective on the simple yet captivating subject of a plate of figs.~~



Detailed: Minimalist still life composition featuring a white rabbit, a blue triangular prism, a pink cylindrical base, and a small bird. **The rabbit is positioned in the lower left foreground**, sitting on a reflective surface, facing right. **The blue prism is centrally placed, balanced on the pink cylinder**. A small bird with a yellow and black head perches on the right edge of the blue prism. In the background, a large green cone stands to the right, partially out of focus. The backdrop is a soft gradient of pink and white, creating a serene and balanced atmosphere. High-key lighting, clean lines, smooth textures, pastel color palette, modern and abstract, visually balanced, playful yet sophisticated.

Medium: A minimalist still life with a white rabbit on a reflective surface, a blue triangular prism on a pink cylinder, and a small bird perched on the prism. A large green cone stands in the background. The scene features a pastel color palette, high contrast, and a clean, modern aesthetic with a balanced composition.

Short: A minimalist still life featuring a blue triangular prism on a pink cylinder with a perched bird and a white rabbit, set against a pastel background with a green cone, creating a balanced and whimsical composition.

Tag: 3D rendering, surreal composition, geometric shapes, pastel colors, minimalist design, whimsical, high contrast, glossy textures, playful, visually balanced.

Detailed: 超现实主义的静物构图，以几何图形和动物为主题，背景为柔和的色彩。**一只白色兔子坐在左前景**，面向右侧，耳朵竖起。**兔子左侧是一个粉色圆柱形底座，支撑着一个巨大的蓝色三角形**。一只小而多彩的鸟儿，羽毛为黄色和黑色，**停在三角形的右边缘**。背景中有一个绿色的圆锥体直立着，部分可见于右侧。背景由柔和的粉红色和白色垂直条纹组成。营造出一种诙谐和极简的氛围。高调照明，柔和阴影，鲜艳的颜色搭配，光滑的质感，现代艺术风格，平衡的构图，既有趣又宁静的氛围。

Medium: 一个极简主义的3D场景展示了一只白色兔子坐在粉色圆柱上，旁边是一个蓝色三角形棱镜，上面停着一只小黄鸟。背景中有一个绿色圆锥体，背景是柔和的粉红色和白色渐变。设计强调干净的线条、鲜艳的色彩和现代、有趣的审美。

Short: 一个极简主义的3D渲染场景，展示了一个蓝色三角形棱镜置于粉色圆柱体上，一只黄色和黑色的鸟儿停歇在棱镜之上，一只白色兔子坐在前景，背景是柔和的粉红色和白色，营造出一个充满活力、超现实且和谐的构图。

Tag: 极简主义，几何形状，鲜艳色彩，光滑质感，现代艺术，奇幻，鸟，兔子，三角形，圆柱形，绿锥体，柔和灯光，高对比度，抽象，当代艺术

Short (Florence): The image shows a white rabbit sitting on top of a pink and blue pyramid-shaped object, with a bird perched on top. To the right of the rabbit is a green object, and in the background is a wall.

Detailed: In the image, a white rabbit is sitting on a pink pedestal. The rabbit is positioned on the left side of the pedestal. On the right side of the pedestal, there is a bird perched on a blue triangle. The bird is facing the rabbit, creating a sense of interaction between the two animals. In the background, there is a green cone ~~and a purple cylinder~~, adding a touch of color to the scene. The objects are arranged in a way that draws attention to the interaction between the rabbit and the bird, making it the focal point of the image. The overall scene is a blend of colors and shapes, creating a visually appealing composition.

Figure 8: Comparison with ShareGPT4V [46] and Florence [62] in visual understanding and spatial relationships. The **blue underline** correspond to areas with more detailed and accurate descriptions, while **red underline** and **red strikethrough** represent the incorrect and insufficient descriptions respectively.

and loss curve trends. In the high-quality tuning stage, the model achieves substantial improvements within just 1K steps, e.g. from 85.7 to 86.6 on the DPG benchmark and from 0.67 to 0.71 on the GenEval. However, as high-quality tuning progressed, performance fluctuations are observed. Notably, at 11K steps in the third stage, the model continues improving on DPG benchmark, whereas the performance declines slightly on GenEval.

Ablation Study on Efficient Inference Strategy. In Fig. 11, we evaluate the inference efficiency of Lumina-Image 2.0 under a 1024-resolution setting using multiple inference strategies, including CFG-Renorm, CFG-Trunc, FDPM, and TeaCache. First, our results show that the proposed CFG-Renorm and CFG-Trunc fusion method (Sec. 4.5) not only saves sampling time but also has a negligible impact on the quality of the sampled results. Second, integrating FDPM into our model can effectively reduce inference time. However, empirical evaluations indicate that FDPM suffers from poor stability, negatively affecting sample quality during the generation process. Third, while incorporating TeaCache further improves sampling speed, it significantly degrades image quality, often leading to blurriness. As a result, in



Figure 9: High-quality image generation examples from Lumina-Image 2.0, showcasing its precise prompt-following ability and its capability to generate highly aesthetic and realistic images across different resolutions.

Table 8: Performance Comparison Across Stages on DPG [30] and GenEval [31] Benchmarks.

Stage	Steps (K)	DPG	GenEval
Low Res. Stage	15	84.5	0.63
High Res. Stage	38	85.7	0.67
HQ Tuning Stage	1	86.6	0.71
HQ Tuning Stage	5	87.2	0.73
HQ Tuning Stage	11	87.6	0.72

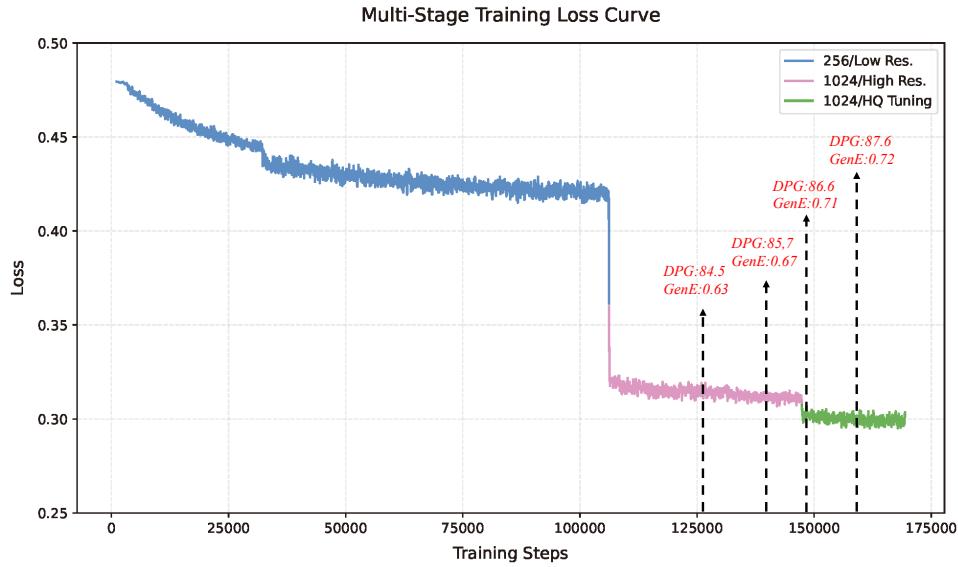


Figure 10: Loss curves for the three training stages, showing a steady performance increase in the DPG [30] and GenEval [31] benchmark.

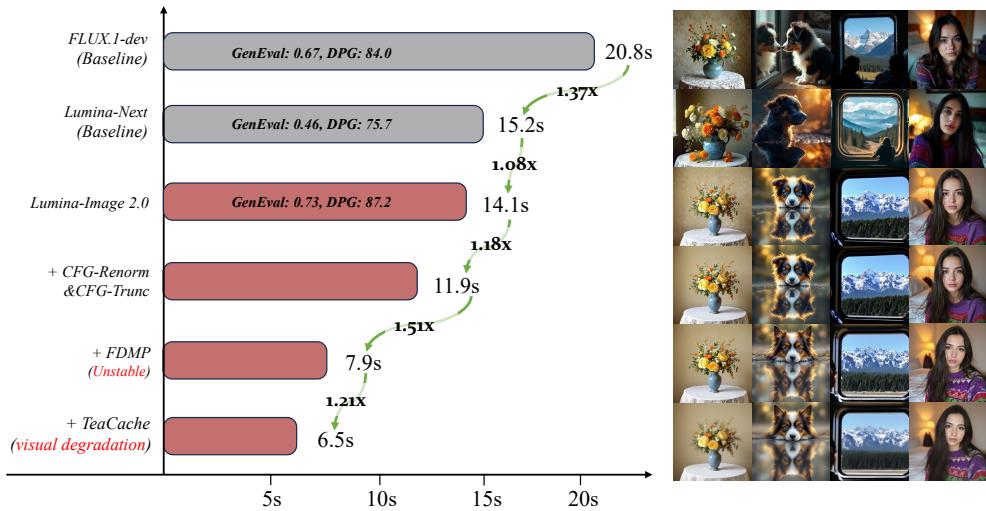


Figure 11: Ablation study on efficient inference strategy. The performances are measured on a single A100 GPU with batch size 1.

practical applications, we adopt Lumina-Image 2.0 with CFG-Renorm and CFG-Trunc as the final solution to balance efficiency and quality.



Figure 12: Generation defects of Lumina-Image 2.0, categorized into overall structural errors, texture detail errors, and text errors.

6 Limitation

Although we have followed previous works [8, 15, 44, 10, 37] to evaluate our method on benchmarks such as GenEval [31] and T2ICompBench [32], achieving comparable performance with state-of-the-art models, we argue that these academic benchmarks are not comprehensive and may sometimes fail to accurately assess image quality in alignment with human perception. To illustrate this point, Fig. 12 highlights several limitations of Lumina-Image 2.0. First, for complex and diverse structures (e.g., human bodies) and for rare concepts in the training data (e.g., handguns), our model struggles to consistently generate correct results. Second, when handling images with intricate textures, such as densely crowded scenes or tire spokes, our model frequently generates disordered details. Finally, our model still needs substantial improvements in accurately rendering long and complex text.

7 Conclusion

This paper introduces Lumina-Image 2.0, a unified and efficient text-to-image generative framework that achieves strong performance in both image quality and prompt alignment. Specifically, a Unified Next-DiT model is developed to generate high-quality images through the seamless integration of textual and visual information. A Unified Captioner (UniCap) is proposed to produce detailed and accurate textual descriptions for constructing high-quality image-text training pairs. In addition, a set of efficient training and inference strategies is developed to further optimize performance while reducing computational costs. Lumina-Image 2.0 achieves promising performance on public benchmarks, and provides a transparent, reproducible text-to-image generative framework. We hope that our model will contribute to advancing the field of text-to-image generation.

References

- [1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.
- [2] 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.
- [3] 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.
- [4] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

- [5] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. [arXiv preprint arXiv:2405.09818](#), 2024.
- [6] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. [arXiv preprint arXiv:2410.10812](#), 2024.
- [7] 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.
- [8] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. [Proceedings of the International Conference on Learning Representations \(ICLR\)](#), 2025.
- [9] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2023.
- [10] 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, 2025.
- [11] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In [Proceedings of the IEEE International Conference on Computer Vision \(ICCV\)](#), pages 3836–3847, 2023.
- [12] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In [Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition \(CVPR\)](#), pages 8640–8650, 2024.
- [13] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In [Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition \(CVPR\)](#), pages 6007–6017, 2023.
- [14] Weifeng Lin, Xinyu Wei, Renrui Zhang, Le Zhuo, Shitian Zhao, Siyuan Huang, Junlin Xie, Yu Qiao, Peng Gao, and Hongsheng Li. Pixwizard: Versatile image-to-image visual assistant with open-language instructions. [arXiv preprint arXiv:2409.15278](#), 2024.
- [15] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. [Proceedings of the International Conference on Learning Representations \(ICLR\)](#), 2023.
- [16] 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.
- [17] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. [arXiv preprint arXiv:2406.18583](#), 2024.
- [18] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. [arXiv preprint arXiv:2502.10248](#), 2025.
- [19] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. [arXiv preprint arXiv:2502.04896](#), 2025.
- [20] Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models. [arXiv preprint arXiv:2406.11831](#), 2024.
- [21] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. [arXiv preprint arXiv:2308.06721](#), 2023.
- [22] 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.
- [23] OpenAI. Chatgpt: Optimizing language models for dialogue, 2022.
- [24] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. [arXiv preprint arXiv:2412.16720](#), 2024.
- [25] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. [arXiv preprint arXiv:2407.21783](#), 2024.

- [26] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. [arXiv preprint arXiv:2307.09288](#), 2023.
- [27] Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenzhe Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, et al. Stiv: Scalable text and image conditioned video generation. [arXiv preprint arXiv:2412.07730](#), 2024.
- [28] Mingyang Yi, Aoxue Li, Yi Xin, and Zhenguo Li. Towards understanding the working mechanism of text-to-image diffusion model. [Advances in Neural Information Processing Systems \(NeurIPS\)](#), 2024.
- [29] Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. [arXiv preprint arXiv:2411.19108](#), 2024.
- [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] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. [Advances in Neural Information Processing Systems \(NeurIPS\)](#), 36, 2024.
- [32] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhengu Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. [Advances in Neural Information Processing Systems \(NeurIPS\)](#), 36:78723–78747, 2023.
- [33] Dustin Podell, Zion English, Kyle Lace, 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.
- [34] William S Peebles and Saining Xie. Scalable diffusion models with transformers. In [Proceedings of the IEEE International Conference on Computer Vision \(ICCV\)](#), volume 4172, 2022.
- [35] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhengu Li. Pixart- σ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation. [Proceedings of the European Conference on Computer Vision \(ECCV\)](#), 2024.
- [36] 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.
- [37] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuteng Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. [arXiv preprint arXiv:2409.11340](#), 2024.
- [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In [International conference on machine learning](#), pages 8748–8763. PMLR, 2021.
- [39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. [Journal of Machine Learning Research](#), 21(140):1–67, 2020.
- [40] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. [arXiv preprint arXiv:2403.08295](#), 2024.
- [41] 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.
- [42] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. [arXiv preprint arXiv:2210.02747](#), 2022.
- [43] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. [arXiv preprint arXiv:2408.02657](#), 2024.
- [44] 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. [arXiv preprint arXiv:2412.04431](#), 2024.
- [45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. [Advances in neural information processing systems](#), 36, 2024.

- [46] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024.
- [47] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.
- [48] Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024.
- [49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [50] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvilm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.
- [51] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2, 2023.
- [52] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [54] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023.
- [55] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: enhanced transformer with rotary position embedding. *arxiv*. *arXiv preprint arXiv:2104.09864*, 2021.
- [56] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [57] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [58] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuandvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- [59] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [60] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.
- [61] Kaiyan Zhang, Biqing Qi, and Bowen Zhou. Towards building specialized generalist ai with system 1 and system 2 fusion. *arXiv preprint arXiv:2407.08642*, 2024.
- [62] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.
- [63] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

- [64] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. [arXiv preprint arXiv:2111.03930](#), 2021.
- [65] Sainbayar Sukhbaatar, Edouard Grave, Guillaume Lample, Herve Jegou, and Armand Joulin. Augmenting self-attention with persistent memory. [arXiv preprint arXiv:1907.01470](#), 2019.
- [66] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. [arXiv preprint arXiv:2501.09732](#), 2025.
- [67] 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. [arXiv preprint arXiv:2501.18427](#), 2025.
- [68] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. [arXiv preprint arXiv:2408.03314](#), 2024.
- [69] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. [Advances in Neural Information Processing Systems Workshops \(NeurIPS Workshops\)](#), 2021.
- [70] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. [arXiv preprint arXiv:2211.01095](#), 2022.
- [71] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. [Advances in Neural Information Processing Systems](#), 36:36652–36663, 2023.
- [72] Biao Zhang and Rico Sennrich. Root mean square layer normalization. [Advances in Neural Information Processing Systems](#), 32, 2019.
- [73] I Loshchilov. Decoupled weight decay regularization. [arXiv preprint arXiv:1711.05101](#), 2017.
- [74] 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.
- [75] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, et al. Imagen 3. [arXiv preprint arXiv:2408.07009](#), 2024.
- [76] Arpad E Elo and Sam Sloan. The rating of chessplayers: Past and present. [Arco Pub](#), 1978.
- [77] Kuaishou Technology. Kolors. <https://github.com/Kwai-Kolors/Kolors>, 2024.
- [78] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. [Advances in neural information processing systems](#), 35:36479–36494, 2022.
- [79] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. [arXiv preprint arXiv:2408.00118](#), 2024.