Illuminating Unified Multimodal Model for Free-form Interleaved Text-Image Generation

Chonghuinan Wang[1,*], Zhikai Chen[2,*], Chunwei Wang[2,*], Yecong Wan[1,3], Junwei Yang[2], Zhixin Wang[2], Wei Zhang[2], Jiaqi Xu[2], Renjing Pei[2], Xiaohe Wu[1,✉], Fan Li[2,4,†,✉], and Wangmeng Zuo[1]

[1] Harbin Institute of Technology, Harbin, China
[2] Huawei Noah's Ark Lab, Shenzhen, China
[3] Zhengzhou Advanced Research Institute of Harbin Institute of Technology, Zhengzhou, China
[4] Nankai University, Tianjin, China
25b903050@stu.hit.edu.cn, csxhwu@gmail.com, lifan61@huawei.com
https://github.com/ChonghuinanWang/ILLUME-X

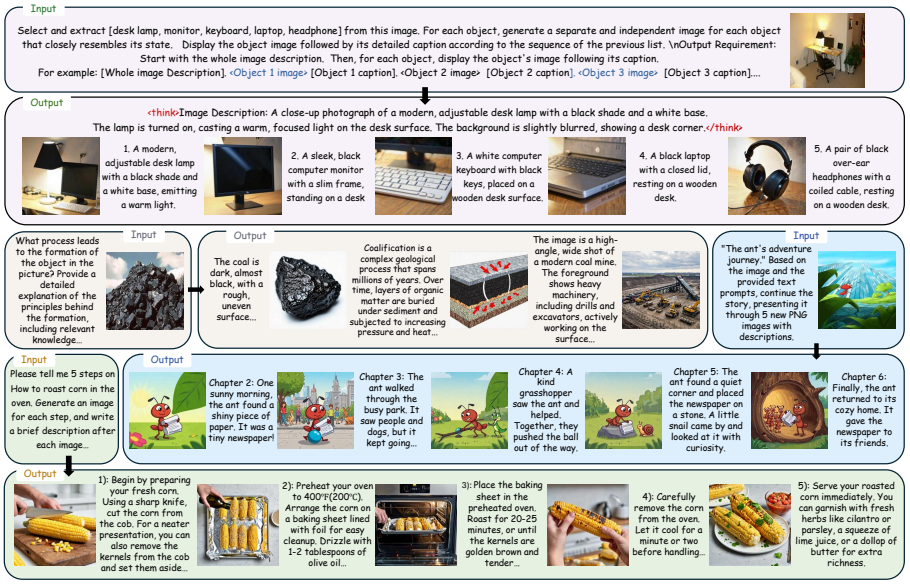

Fig. 1: Illustrative examples of ILLUME-X. The model handles interleaved text-image inputs and outputs, enabling cohesive multimodal understanding and generation.

Abstract. The advancement of generative AI models capable of producing text and image marks a critical step forward in the realm of

* Equal Contribution, ✉ Corresponding Author, † Project Leader

multimodal intelligence, particularly for tasks involving the interleaving of both modalities. To advance this intelligence to the next stage, it is crucial for models to autonomously generate free-form interleaved text-image sequences. In this paper, we introduce **ILLUME-X**, an advanced unified multimodal paradigm that enables high-quality, free-form interleaved text-image generation by improving multimodal data efficiency and stabilizing the multimodal training process. ILLUME-X comprises three key components: (i) an expanded training data pipeline optimized for interleaved text-image generation, (ii) a progressive training strategy with self-adaptive objectives for free-length multimodal token sequences, and (iii) an objective and comprehensive evaluation method **ILScore** for interleaved text-image sequences. Notably, our ILLUME-X outperforms previous unified models across multiple interleaved text-image generation tasks like style transfer, image decomposition and storytelling.

Keywords: Interleaved Text-Image Generation · Multimodal Unified Model · Diffusion Transformer

1 Introduction

Multimodal Large Language Models (MLLMs) and diffusion models [8,12,14,20, 22,34,35,43,47,50,62] have recently achieved significant advances in visual understanding and generation. By merging understanding and generation capabilities within a single framework, unified multimodal models can genuinely grasp the relationships between visual and textual information to enable more intelligent interactions and task execution. Naturally, researchers demonstrate considerable interest in their abilities for interleaved text-image generation, which requires the models to generate both images and text pieces in an arbitrary sequence for diverse real-world scenarios.

To enable multimodal models to perform interleaved generation, existing studies have explored several complementary paradigms. One line of work [12,15, 60] directly applies a pure auto-regression model to generate texts and images. These models utilize a unified auto-regressive Transformer to learn the joint probability distribution of text and visual tokens, and elicit emergent cross-modal generative capabilities by scaling up dataset size and model capacity. Another direction [24,46] attempts to combine the diffusion model into an auto-regression model. These approaches incorporate modality-aware architectural designs within the transformer to support joint prediction of text and image tokens, thereby enabling cross-modal generation in a single framework. However, despite these advancements, existing methods have not fully evaluated the performance of their models or sufficiently validated their capabilities on interleaved text-image generation benchmarks. This gap in comprehensive evaluation and benchmark performance is a key motivation for this work.

In this work, we propose ILLUME-X, a unified multimodal paradigm that enables high-quality, free-form interleaved text-image generation. To empower the model with high-quality interleaved capabilities, we first establish a systematic

data pipeline that curates 100K high-fidelity training samples. This pipeline integrates authentic interleaved sequences derived from video data with sophisticated synthetic data generated by multimodal LLMs, capturing real-world temporal dynamics and diverse generative scenarios. Building upon this data foundation, ILLUME-X employs an integrated architecture that aligns cross-modal representations using shared attention mechanisms. Additionally, an interleaved training paradigm is introduced, featuring a specialized attention mask to jointly optimize text generation, image synthesis, and modality transitions. This enables flexible multi-input-to-multi-output (N-to-M) reasoning. Finally, we propose a rigorous evaluation protocol ILScore to address the lack of task-specific metrics for interleaved generation. This framework evaluates the generated sequences from two critical perspectives: cross-modal continuity (ensuring seamless integration) and unimodal generative quality (ensuring both textual coherence and visual fidelity).

The main contributions of ILLUME-X are as follows:

- It introduces a pioneering unified multimodal paradigm that achieves free-form N-to-M interleaved generation, powered by a progressive training strategy with self-adaptive objectives for free-length multimodal token sequences.
- It presents an efficient and scalable data pipeline that curates 100K high-quality interleaved samples, enhancing model training.
- It establishes a multi-dimensional evaluation protocol ILScore specifically designed for complex text-image sequences, ensuring thorough performance assessment across modalities.
- Extensive experiments demonstrate that ILLUME-X outperforms previous methods, enabling unified MLLMs to generate free-form interleaved content.

2 Related Work

2.1 Unified Multimodal Models

Unified multimodal models have emerged as a key direction for bridging the gap between multimodal understanding and generation, aiming to consolidate diverse tasks (*e.g.* vision-language reasoning, text-to-image synthesis, and image editing) within a single architectural framework. Existing unified models can be broadly grouped into three main categories: diffusion models, autoregressive (AR) models, and hybrid combinations of the two paradigms. Diffusion-based unified models (*e.g.* Dual Diffusion [29], MMaDA [55], LaViDa-O [28], UniModel [59], Muddit [40]) usually adapt text-to-image diffusion pipelines to multimodal scenarios by conditioning the denoising process on cross-modal contextual signals. They offer high visual fidelity but are limited by slow inference and weaker sequential reasoning. AR-based unified models (*e.g.* Chameleon [43], Janus-Pro [8], UniToken [23], Emu 3 [49]) typically serialize vision and language tokens into a single sequence and model them in an ordered, token-by-token manner. AR + diffusion hybrid models (*e.g.* Transfusion [62], Show-o [53], BLIP3-o [6], OmniGen2 [52], Mogao [30], EMMA [20], BAGEL [14], HBridge [48], HunyuanImage [4]) offer a

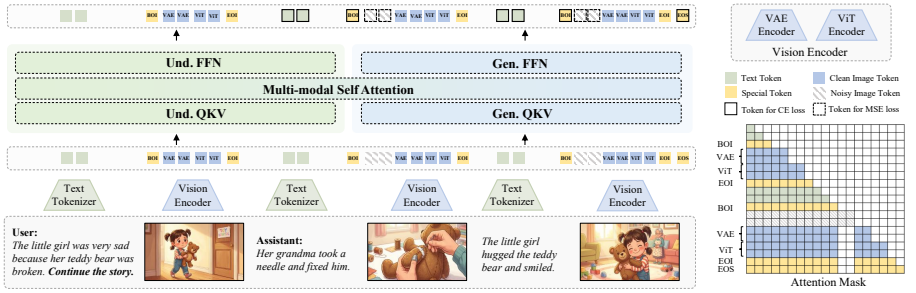

Fig. 2: The overall architecture of our ILLUME-X. "QKV" and "FFN" represent query, key, value vectors for attention computation and feedforward networks, respectively.

compelling compromise, combining the compositional control of autoregressive decoding with the high-fidelity synthesis of diffusion.

2.2 Interleaved Generation Models and Datasets

Although existing unified models can produce both images and text, they are typically constrained to a single target modality specified in advance. While some studies (*e.g.* [4, 12, 14, 27, 30, 51, 54]) appear to demonstrate the feasibility of interleaved image–text generation, dedicated investigation and validation for N-to-M interleaved generation scenarios remain limited. To this end, CoMM [7] aggregates multi-source data with an emphasis on instructional content and visual storytelling to strengthen MLLMs' in-context learning and interleaved image–text generation. In parallel, Zebra-CoT [26] provides large-scale interleaved multimodal reasoning data, and fine-tuning BAGEL [14] on it equips the model with interleaved multimodal reasoning capabilities. However, despite their utility, both datasets remain limited in overall quality and consistency, which may be hinder stable training for N-to-M interleaved generation. Furthermore, frameworks such as WEAVE [10] and ISG-Bench [5] provide a critical foundation for evaluating interleaved generation by establishing standardized task settings and unified evaluation protocols.

3 Approach

3.1 Model Architecture

Unified Transformer. The architecture of ILLUME-X is illustrated in Figure 2. Following established practices [8,50], we employ distinct continuous visual encoders to support both visual understanding and generation tasks. Specifically, a ViT encoder is used to extract visual semantic features, while a VAE encoder captures low-level visual details. The backbone model adopts a decoder-only

transformer architecture, consistent with modern large language models [2, 3]. Following the MoT [14] paradigm, our transformer processes token sequences via shared self-attention across all layers, and employs modality-specific QKV projections and FFNs to support both image understanding and generation. To ensure stable and efficient training, the model integrates several advanced components, including RMSNorm [58] for normalization, QK-Norm [13,17] to stabilize attention gradients, SwiGLU [39] as the activation function, and RoPE [42] for token positional encoding. Additionally, we utilize Grouped-Query Attention [1] to reduce KV cache overhead and improve inference efficiency.

Interleaved Attention. To handle free-form interleaved multimodal contexts, our framework concatenates text tokens, vision tokens and special tokens into a unified sequence. Figure 2 illustrates the token organization mechanism and their attention mask in our model. Besides the tokens encoded from text and image, we introduce special tokens, including BOI (begin-of-image), EOI (end-of-image), and EOS (end-of-sequence) tokens, to explicitly demarcate modality transitions. Consequently, text and vision tokens from both understanding and generation tasks are interleaved within the model according to the structural layout of the input. For tokens within the same sample, we employ a generalized causal attention mechanism. Specifically, tokens are partitioned into consecutive segments based on their modality. While tokens in any given segment can attend to all tokens in preceding segments, the attention within each segment is task-specific: we apply causal attention to text and special tokens to maintain the auto-regressive property, while utilizing bidirectional attention for vision tokens to capture holistic spatial information. We explicitly arrange the noisy latent tokens and clean reconstruction tokens within the VAE latent sequence. To alleviate the error accumulation caused by noise interference, we apply a masking operation to the noisy image tokens during the generation of the next set of samples.

Training Objective. During training, we employ a multi-objective optimization strategy to jointly supervise the interleaved sequences. Specifically, a Cross-Entropy (CE) loss is applied to text and special control tokens (BOI/EOS) to maintain language modeling proficiency and ensure seamless modality transitions. Concurrently, vision tokens are optimized via a Rectified Flow-based mean squared error (MSE) loss [17, 33], enabling high-fidelity image synthesis within the unified multimodal stream.

3.2 Interleaved Classifier-free Guidance

For semantic alignment control in image generation, we use Classifier-Free Guidance (CFG) to enhance generation quality, which combines conditional predictions with unconditional predictions to yield results that more closely align with the specified conditions, formulated as:

$$\nabla_x \log p(x|c) = \gamma(\nabla_x \log p(x|c) - \nabla_x \log p(x)) + \nabla_x \log p(x), \tag{1}$$

where c and γ denote the condition and CFG coefficient, respectively. Unlike the condition of image understanding and generation models, which consists only of

the text modality, the condition c and the output x in the interleaved multimodal generation scenario contains a sequence of texts and images, defined as:

$$c = \{c^0_{txt}, c^0_{img}, c^1_{txt}, c^1_{img}, \dots\}, \quad x = \{x^0_{txt}, x^0_{img}, x^1_{txt}, x^1_{img}, \dots\}. \tag{2}$$

Previous models [14] leverage classifier-free guidance with respect to both text and image conditions, and randomly drop each c^j_{txt} or c^j_{img} independently during training. To address this complexity, we decompose the conditions into $c_{txt} = \{c^j_{txt}\}$, $c_{img} = \{c^j_{img}\}$ and the targets into $x_{txt} = \{x^j_{txt}\}$, $x_{img} = \{x^j_{img}\}$ according to their modalities, and then use an interleaved classifier-free guidance mechanism on image generation [30], which is formulated as:

$$\begin{aligned} \nabla_{x_{img}} \log p(x_{img}|c_{txt}, c_{img}) = \gamma_{txt}(&\nabla_{x_{img}} \log p(x_{img}|c_{txt}, c_{img}) \\ &- \nabla_{x_{img}} \log p(x_{img}|c_{img})) \\ &+ \gamma_{img}(\nabla_{x_{img}} \log p(x_{img}|c_{img}) \\ &- \nabla_{x_{img}} \log p(x_{img})) + \nabla_{x_{img}} \log p(x_{img}). \end{aligned} \tag{3}$$

In this way, we simplify the two-stage guidance formulation, with text-condition CFG as the core anchor and separate scaling coefficients $\gamma_{txt}, \gamma_{img}$ for image conditions. During training, we randomly take three condition sampling strategies on the conditions: keep both text and image conditions, only drop the text condition, and drop both text and image conditions simultaneously. Crucially, to maintain the model's ability to process the special tokens in the context, we exclude losses for text tokens and special tokens when we drop any conditions. As such, we can adjust the text guidance scale γ_{txt} and image guidance scale γ_{img} in inference to enhance image generation quality.

3.3 Training Strategy

Pre-trained unified MLLMs are typically optimized across multimodal understanding and generation tasks. To adapt pretrained unified MLLMs for diverse interleaved text-image tasks, we use a supervised training pipeline with two complementary data components under a unified training framework.

In-context Generation Training. This data component centers on in-context image generation samples. It enhances the model's generative capabilities, which in turn alleviates the optimization challenges typically encountered during interleaved multimodal training. Moreover, it effectively addresses the performance stagnation when generating multiple sequential images by ensuring the model maintains dynamic generation quality throughout the process.

Interleaved Text-Image Training. We further incorporate a mixed corpus for interleaved text-image joint training, encompassing data of image understanding, image generation, and various interleaved text-image tasks. This hybrid training approach enables the model to achieve a comprehensive balance across different objectives. By training on these heterogeneous datasets simultaneously, the model can effectively strengthen its cross-task generalization and improve its performance across a wide range of tasks.

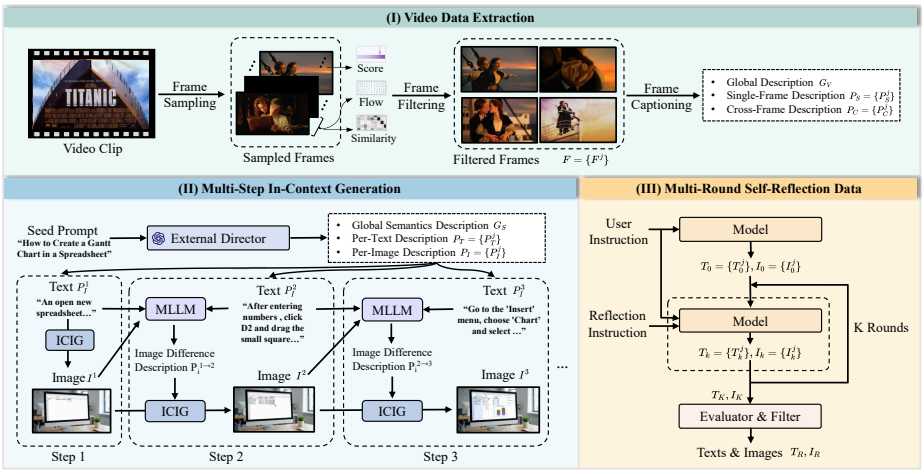

Fig. 3: The overall architecture of our data pipeline. "MLLM" and "ICIG" represent multimodal large language model and in-context image generator, respectively, used for generating textual descriptions and images.

4 Interleaved Data Curation

For in-context generation training, we use datasets established in previous research [14, 52]. Regarding interleaved text-image training, we aggregate existing data from VINCIE [38], SEED-Story [56], and CoMM [7]. However, these open-source resources frequently exhibit limitations such as mediocre image quality, constrained instruction fidelity, and lack of task diversity. To address these shortcomings and align with our research objectives, we have curated an additional new comprehensive training dataset. The following sections elaborate on our data construction via three methods.

4.1 Video Data Extraction

Videos inherently encompass diverse visual information, including variations in character actions, object states, and scene viewpoints. Effective construction of high-quality interleaved text and image data requires the extraction of representative frames that capture major semantic transitions in chronological order and providing their associated textual descriptions rich in causal logic. Our approach addresses this need by developing a multi-stage video data extraction pipeline comprising three main procedures: frame sampling, frame filtering, and frame captioning. Figure 3 (I) illustrates this procedure to extract data from videos.

Frame Sampling. Common frame sampling strategies, such as key-frame and uniform sampling [52], often struggle to align with actual visual content. Specifically, key-frames may not accurately reflect semantic shifts, resulting in uneven sample distributions or inconsistent semantic gaps. Furthermore, standard uniform sampling struggles with varying content velocities across different

video segments, which may lead to significant information loss in dynamic sections. Considering these limitations, we propose a hybrid approach that implement a multi-interval sliding window sampling strategy to capture visual dynamics at various temporal scales to ensure both representativeness and effectiveness of the sampled frames. In practice, we uniformly sample adjacent k-frame windows (e.g., 2, 5 or 8 frames) at varying intervals (e.g., every 1, 3, or 10 seconds), and subsequently refine this selection through a multi-perspective filtering process. This strategy effectively captures the most informative frames while maintaining a coherent temporal flow.

Frame Filtering. To ensure high visual quality and minimize redundancy between adjacent frames, we implement a dual-stage content-based filtering process focusing on aesthetic quality and motion dynamics. (i) Aesthetic Filtering: To prioritize visual quality, we calculate each frame's Laplacian variance and the MANIQA [57] metric as aesthetic score. Frames falling below a predefined score threshold are discarded as low-quality samples. To align with human preferences and ensure semantic coherence, we further utilize Qwen-3-VL-32B [2] as a high-level evaluator to determine whether each frame should be discarded or preserved based on its content integrity. (ii) Motion Filtering: To eliminate the redundant frames, we use RAFT [45] to estimate the optical flow between the source frame and target frame from given adjacent frames. Based on the computed flow fields, we exclude the target frames exhibiting either low mean or low variance, which typically indicates static content or signifies uniform global motion (e.g., camera translation or jitter) rather than meaningful object-level dynamics. To further reduce redundancy, we utilize DINOv2 [36] to extract robust feature representations from the adjacent frame pairs and evaluate their feature differences, thus frames with negligible differences are filtered out as redundant. In this way, we obtain multiple images $F = \{F^j\}$ for one piece of data.

Frame Captioning. To achieve comprehensive frame captioning, we propose a multi-level descriptive method that integrates intra-frame description with inter-frame analysis. Specifically, we leverage the MLLM Qwen-3-VL-32B [2] to generate descriptions across three granularities: (i) Global Descriptions: These aim to synthesize the overarching narrative and thematic blueprint from the sequence of consecutive frames, which goes beyond literal visual mapping, instead emphasizing causal reasoning, temporal coherence, and the logical completeness of the depicted actions. (ii) Single-frame Descriptions: These capture static visual attributes, including meticulous details such as color, quantity, morphology, and the spatial arrangement of objects relative to their background. (iii) Cross-frame Descriptions: These focus on intricate inter-frame dynamics, encompassing: (1) subject-level dynamics, such as shifts in posture, orientation, facial expressions, and motion magnitude; (2) interaction states, involving variations in tools, targets, or agents; and (3) environmental nuances, including subtle transitions in lighting, shadows, and spatial structure.

In this way, we collect the frames$\{F^j\}$ and descriptions G_V, P_S, P_C as a term of interleaved text-image data, serving to train the model's capacity to learn the temporal evolution of actions or events.

4.2 Multi-Step In-Context Generation

The advanced image generation model has strong in-context generation abilities. It is able to extract visual concepts, such as objects or identities, from images and reproduce them in new images. This ability allows for the automatic creation of multi-image sequences with internal associations, serving as interleaved image-text data. To construct this dataset, we develop a multi-step collaborative workflow. This workflow integrates an external MLLM as a director to implement chain-of-thought (CoT) reasoning and imagination for analytical understanding, and adopts an in-context image generator to achieve high-quality final image synthesis.

Figure 3 (II) illustrates the multi-step generation pipeline involves a CoT-based textual planning phase followed by iterative synthesis. First, an external MLLM (*e.g.*, GPT-5 [41]) creates a textual blueprint includes the global semantics description G_S, per-text description $P_T = \{P_T^j\}$ and per-image description $P_I = \{P_I^j\}$. Next, the image sequence is generated iteratively based on these descriptions. The first image I^1 is generated directly from its text prompt P_I^1. After that, subsequent images are produced through a feedback loop. Specifically, we employ Qwen3-VL-32B [2] as another MLLM to analyze and generate image-to-image semantic differences that describe the visual changes needed from one image I^j to the next I^{j+1}. The image difference description, along with the previous visual context I^j, are fed into the in-context image generator (*e.g.*, Gemini 3 Pro [11]) to guide the precise image synthesis, resulting in the next image I^{j+1}.

Finally, this process creates interleaved data that includes global description G_S, texts $P_T = \{P_T^j\}$, images $I = \{I^j\}$, and texts $P_I = \{P_I^j\}$. Trained on such curated data, models are able to capture the chain-of-thought logic underlying the evolution of text-image narratives and corresponding visual variations.

4.3 Multi-Round Self-Reflection Data

Inspired by recent advances in test-time scaling [31] and self-reflection for unified multimodal models [52], the paradigms are extended to interleaved multimodal generation to explore how iterative refinement can build text and image sequence.

To construct the self-reflection dataset, the process is initiated by generating initial texts T_0 and images I_0 from curated user instructions. A powerful external MLLM (e.g., Gemini 3 Pro [11]) then acts as a critic, evaluating these outputs against the original requirements. Upon detecting instruction misalignment or aesthetic deficiencies, the critic diagnoses specific flaws and formulates rectification guidance. This feedback is then prepended to the original user prompt to steer the model through rounds of iterative refinement, culminating in the final reflective data T_K, I_K. Finally, these data are further evaluated by the external multimodal model to discard the suboptimal samples, ensuring high data fidelity and yielding T_R and I_R.

5 Experiments

In this section, we conduct extensive experiments to validate the effectiveness of ILLUME-X as a unified model across a diverse set of vision–language tasks. We primarily evaluate the model on interleaved text-image generation and further benchmark its performance on standard text-to-image generation tasks.

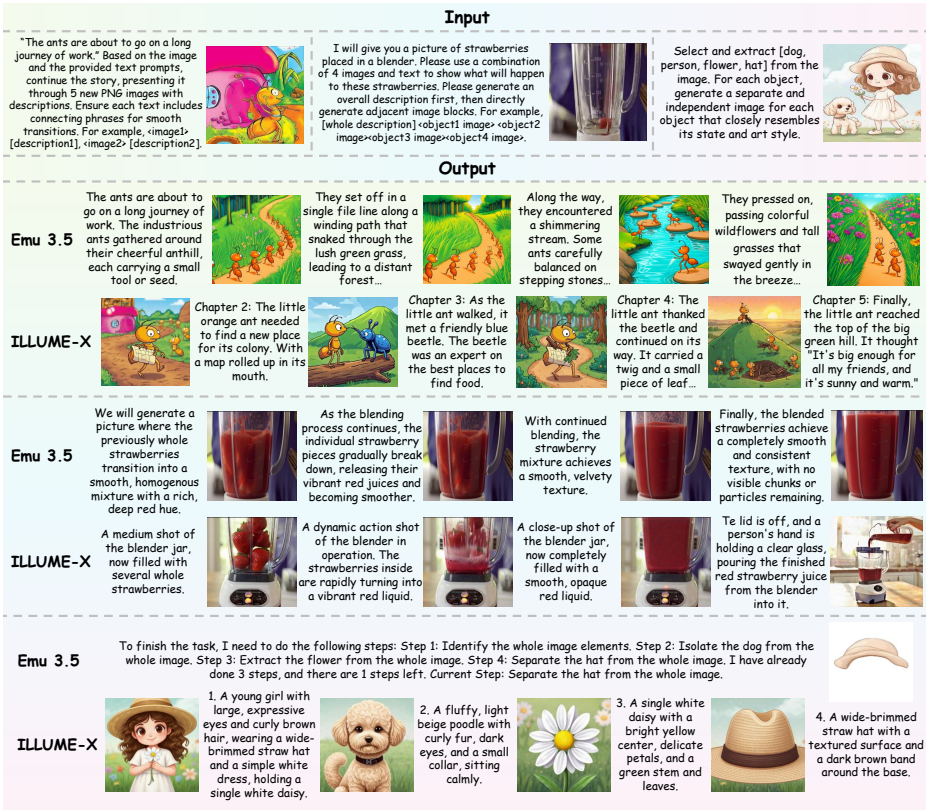

Fig. 4: Qualitative comparison with other methods. We only show the first 4 groups when more than 4 generations are available.

5.1 Evaluation

The widely adopted benchmark for interleaved text-image generation, ISG-Bench [5], has notable evaluation limitations. Several of its metrics heavily depend on the structure of model outputs. As a result, when the structure of the generated results deviates from the predefined template, the evaluation mechanisms tend to fail. This makes it challenging to obtain an objective and robust measurement

of a model's interleaved text-image generation capability. To address these issues, we systematically extend and refine its evaluation dimensions and propose a novel metric termed ILScore.

ILScore. We present a hierarchical quantitative evaluation metric for interleaved text-image generation, which assesses both the global and local quality of generated text-image content from four distinct dimensions: 1) **image-text accuracy**, which measures overall text-image alignment in terms of coherence, relevance, and logical consistency; 2) **single image accuracy**, which evaluates each image based on concept consistency, relationship matching, detail fidelity, visual clarity, and aesthetics; 3) **image sequence accuracy**, which assesses cross-image content and stylistic consistency within multi-image sequences; and 4) **text accuracy**, which independently evaluates the quality and expressiveness of text-only outputs. Together, these four dimensions provide complementary and holistic coverage, enabling precise evaluation across overall text-image alignment, single-image quality, multi-image coherence, and text-only accuracy.

5.2 Implementation details

We choose BAGEL [14] as the initialization of our model weights for its superior performance and public availability. The model is trained with open datasets such as SEED-Story [56], VINCIE [38], WEAVE [10], and our 100K curated dataset. The base learning rate is set to 2×10^{-5}.

Table 1: Quantitive comparison across different tasks on ISG-Bench [5].

Model	Is unified model?	Style Transfer	Progressive	3D_Scene	Image Decomposition	Image-Text Complementation	Temporal Prediction	Visual StoryTelling	VQA	**AVG**
Show-o [53]	✓	2.11	2.41	1.43	2.87	2.06	2.58	3.32	1.86	2.33
Anole [9]	✓	2.93	2.76	1.85	1.49	3.21	2.58	2.97	4.70	2.81
Minigpt-5 [61]	✓	2.15	3.15	1.79	2.54	2.72	2.73	2.91	4.29	2.79
CoMM-Minigpt-5 [7]	✓	2.96	2.60	3.09	2.24	3.09	2.52	2.72	2.87	2.96
Seed-Llama-14b [18]	✓	1.84	3.30	1.52	3.69	1.94	1.78	2.84	2.20	2.39
Gemini [44] & SD3 [16]	✗	4.89	**6.59**	2.68	7.26	6.37	5.26	5.68	**7.89**	5.83
ISG-AGENT [5]	✗	5.87	6.46	**4.89**	**7.58**	6.93	4.54	7.03	6.80	6.26
ILLUME-X	✓	**6.02**	5.75	3.96	6.97	**7.06**	**6.43**	**7.25**	6.67	**6.26**

5.3 Main Results

Qualitative Comparison on Interleaved Text-Image Generation. As shown in Figure 4, ILLUME-X significantly excels in narrative consistency and complex instruction following compared to Emu 3.5 [12]. Specifically, in the first group, ILLUME-X generates a more precise and cohesive narrative with a visual representation that seamlessly corresponds to the description of the journey of the ants, showcasing its strength in visual storytelling. Similarly, in the second group, ILLUME-X produces a smooth transition from whole strawberries to blended juice, illustrating its strength in handling complex transformations in image generation. In contrast, Emu 3.5 [12] often struggles with more abstract connections between text and image, resulting in less coherent outputs in some

Table 2: Quantitative evaluation of various models across the four ILScore dimensions. We report performance across eight sub-tasks and calculate the average score for: D1 (image-text accuracy), D2 (single image accuracy), D3 (image sequence accuracy), and D4 (text accuracy). Bold values indicate the best overall performance in each category.

	Model	Style Transfer	Progressive	3D Scene	Image Decomposition	Image-Text Complementation	Temporal Prediction	Visual StoryTelling	VQA	**AVG**
D1	Emu 3.5 [12]	2.16	4.20	2.50	2.67	6.40	5.40	4.42	5.00	4.09
	Gemini 3 pro [11]	4.10	3.93	3.38	1.11	4.00	3.40	4.64	2.00	3.32
	ILLUME-X	3.80	3.60	3.12	4.60	4.56	3.80	6.64	2.40	4.07
D2	Emu 3.5	7.44	7.19	4.80	5.48	7.58	7.73	7.27	5.75	6.66
	Gemini 3 pro	4.42	6.51	5.08	4.50	9.42	7.46	4.00	9.76	6.39
	ILLUME-X	6.33	5.84	5.93	7.21	6.41	5.96	7.44	6.41	6.44
D3	Emu 3.5	5.05	3.93	3.50	4.70	7.00	8.20	4.25	2.30	4.87
	Gemini 3 pro	2.75	6.20	2.38	3.22	3.89	7.30	4.18	4.80	4.34
	ILLUME-X	6.25	4.47	6.12	4.60	5.89	6.00	5.93	5.70	5.62
D4	Emu 3.5	2.70	5.57	5.25	2.40	8.80	8.00	5.58	7.20	5.69
	Gemini 3 pro	6.50	5.33	5.62	2.56	5.44	5.10	6.55	5.20	5.29
	ILLUME-X	4.25	4.53	4.62	5.60	6.33	5.90	7.93	2.80	5.25
Overall	Emu 3.5	4.34	5.22	4.01	3.81	**7.45**	**7.33**	5.38	5.06	5.33
	Gemini 3 pro	4.44	**5.49**	4.12	2.85	5.69	5.82	4.84	**5.44**	4.84
	ILLUME-X	**5.16**	4.61	**4.95**	**5.50**	5.80	5.42	**6.99**	4.33	**5.34**

Table 3: Cost Comparison.

Model	Parameters	Time per Image
Emu 3.5	34B	409.50s
ILLUME-X (Ours)	7B + 7B	81.33s

cases. Overall, ILLUME-X demonstrates superior performance in generating visually aligned and narratively consistent results across different tasks, making it a strong contender in interleaved multimodal generation tasks.

Quantitative Comparison on Text-Image Generation. For the interleaved text-image task, we first follow the popularly used ISG-Bench [5] to quantitatively evaluate interleaved text-image generation. As illustrated in Table 1, our proposed ILLUME-X achieves state-of-the-art (SOTA) performance among all unified models, reaching an average score of 6.26. Notably, ILLUME-X not only outperforms existing unified models like Anole (2.81) and MiniGPT-5 (2.79) by a significant margin (over 120% improvement in AVG score) but also matches the performance of the non-unified agent-based system, ISG-AGENT (6.26). This demonstrates the effectiveness of our approach in handling complex interleaved generation tasks within a single unified framework.

Then, we evaluate our ILLUME-X, open-source Emu 3.5 and commercial models Gemini 3 Pro on our proposed ILScore. As illustrated in Table 2, ILLUME-X overall achieves the highest aggregate score (5.34), marginally surpassing Emu 3.5 (5.33) and significantly outperforming Gemini 3 Pro (4.84). ILLUME-X demonstrates superior robustness in complex structural tasks, securing top honors in four out of eight categories: Style Transfer, 3D Scene, Image Decomposition, and Visual Storytelling. More notably, our method substantially reduces

training overhead and achieves markedly lower inference latency, as validated in Table 3, making it better suited for practical real-world deployment scenarios.

Table 4: Comparison of different models on GenEval [19] benchmark.

Model	# Params	GenEval [19]						
	B	Single object	Two objects	Counting	Colors	Position	Color attribution	Overall
SDXL [37]	6.6	0.98	0.74	0.39	0.85	0.15	0.23	0.55
FLUX.1-dev [25]	8+12	0.99	0.81	0.79	0.74	0.20	0.47	0.67
Show-o [53]	1.3	0.98	0.80	0.66	0.84	0.31	0.50	0.68
Janus-Pro [8]	7	0.99	0.89	0.59	0.90	**0.79**	0.66	0.80
UniWorld-V1 [32]	7+12	0.99	0.93	0.79	0.89	0.49	0.70	0.80
OmniGen2 [52]	3+4	**1.00**	**0.95**	0.64	0.88	0.55	**0.76**	0.80
BAGEL [14]	7+7	0.99	0.94	0.81	0.88	0.64	0.63	0.82
ILLUME-X	7+7	0.99	**0.95**	**0.81**	**0.92**	0.71	0.74	**0.85**

Table 5: Comparison of different models on DPG-Bench [21] benchmark.

Model	# Params	DPG-Bench [21]					
	B	Global	Entity	Attribute	Relation	Other	Overall
SDXL [37]	6.6	83.27	82.43	80.91	86.76	80.41	74.65
FLUX.1-dev [25]	8+12	82.1	89.5	88.7	**91.1**	89.4	84.0
Show-o [53]	1.3	79.33	75.44	78.02	84.45	60.80	67.27
Janus-Pro [8]	7	86.90	88.90	89.40	89.32	89.48	84.19
UniWorld-V1 [32]	7+12	83.64	88.39	88.44	89.27	87.22	81.38
OmniGen2 [52]	3+4	88.81	88.83	90.18	89.37	**90.27**	83.57
BAGEL [14]	7+7	88.94	90.37	91.29	90.82	88.67	85.07
ILLUME-X	7+7	**90.28**	**91.80**	**94.08**	82.37	88.00	**86.38**

More Comparison on the Text-to-Image Task. We also assess the generation quality for the text-to-image task using widely adopted benchmarks, GenEval [19] and DPG-Bench [21], to evaluate both basic semantic alignment and complex scene synthesis. As shown in Table 4, ILLUME-X outperforms other models on the GenEval benchmark, achieving an overall score of 0.85, surpassing OmniGen2 (0.80) and BAGEL (0.82). ILLUME-X excels in Counting (0.81), Colors (0.92), and Overall performance (0.85), while also showing strong results in Single Object (0.99) and Two Objects (0.95). Its performance in Color Attribution (0.74) and Position (0.71) further underscores its effectiveness in handling complex scene synthesis, making it a competitive choice in text-to-image generation.

As shown in Table 5, it presents a comprehensive evaluation of image generation performance on DPG-Bench. ILLUME-X achieves the highest Overall score of 86.38, surpassing strong competitors such as BAGEL (85.07) and Janus-Pro (84.19). A key driver of this performance is our model's exceptional accuracy in Attribute (94.08) and Entity (91.80) categories, where it outperforms all baseline models, including the significantly larger FLUX.1-dev. The substantial margin

in the Attribute score (+2.79 over BAGEL) highlights ILLUME-X's precision in grounding specific textual descriptors to visual entities.

Interestingly, while our model shows a slight performance gap in the Relation sub-category compared to FLUX.1-dev, its Global (90.28) score indicates a superior ability to maintain holistic scene coherence. Combined with its efficient 7B+7B parameter architecture, these results further validate the effectiveness of ILLUME-X in handling dense and complex prompts that require fine-grained semantic alignment.

Table 6: Ablation Results. Quantitative Comparison of different CFG parameters and CoT settings for text-image interleaved generation. γ_{txt} represents the text guidance scale, and γ_{img} denotes the image guidance scale.

γ_{txt}	γ_{img}	w/ CoT	Visual Storytelling	Image-Text Complementation
2.0	1.0	✗	7.02	5.35
4.0	1.0	✗	7.13	5.59
8.0	1.0	✗	7.12	5.69
8.0	1.25	✗	7.00	5.52
8.0	1.5	✗	6.85	5.60
8.0	1.0	✓	**7.28**	**5.91**

5.4 Ablation Study

To efficiently and targetedly validate the effect of the interleaved text-image structure, we conduct ablation studies only on the Visual Storytelling and "howto" subset of Image-Text Complementation. These two task categories inherently rely on temporal logic and strong multimodal complementarity, making them representative scenarios where interleaving is expected to yield the most pronounced gains. All remaining tasks adopt the best-performing structure identified through ablation for fair comparison. Experimental results demonstrate that the best performance is achieved when the CFG uses a text guidance scale of 8.0, an image guidance scale of 1.0, and incorporates CoT. We visualize qualitative examples with and without CoT under these optimal hyperparameters in Figure 5.

6 Limitations

ILLUME-X is primarily trained and evaluated for interleaved text-image generation at a resolution of 512. Due to limitations imposed by the underlying model architecture and the finite context length, scaling both training and inference to resolutions of 1024 and above remains challenging. Consequently, the quality of high-resolution interleaved generation still leaves room for further improvement.

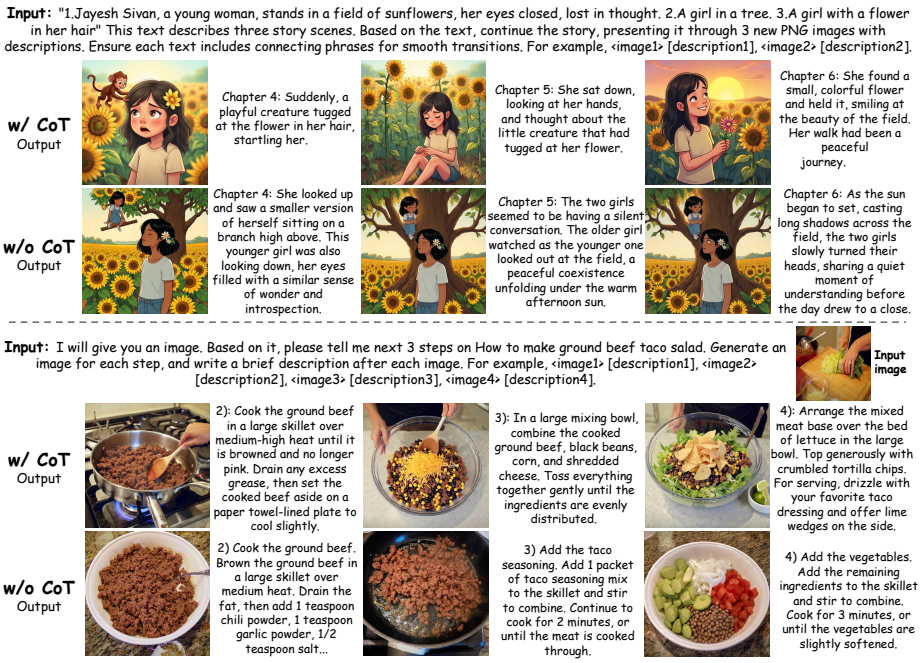

Fig. 5: Qualitative comparison of different CoT settings.

7 Conclusion

In this work, we present ILLUME-X, a unified multimodal model that supports free-form interleaved text-image generation. We also build an efficient multimodal interleaving data construction pipeline to produce training samples. In addition, we extend existing evaluation metrics and introduce ILScore to better reflect the requirements of interleaved text-image generation. Extensive comparative experiments demonstrate the effectiveness of the proposed framework.

8 Acknowledge

This work was supported by the National Cyber Security-National Science and Technology Major Project under Grant No. 2026ZD1500500 and the National Natural Science Foundation of China (NSFC) under Grant No.62476067.

References

1. Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebron, F., Sanghai, S.: GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In: EMNLP (2023)

2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL Technical Report. arXiv 2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv 2502.13923 (2025)
4. Cao, S., Chen, H., Chen, P., Cheng, Y., Cui, Y., Deng, X., Dong, Y., Gong, K., Gu, T., Gu, X., et al.: Hunyuanimage 3.0 technical report. arXiv preprint arXiv:2509.23951 (2025)
5. Chen, D., Chen, R., Pu, S., Liu, Z., Wu, Y., Chen, C., Liu, B., Huang, Y., Wan, Y., Zhou, P., et al.: Interleaved scene graphs for interleaved text-and-image generation assessment. arXiv preprint arXiv:2411.17188 (2024)
6. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
7. Chen, W., Li, L., Yang, Y., Wen, B., Yang, F., Gao, T., Wu, Y., Chen, L.: Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. In: CVPR (2025)
8. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. arXiv 2501.17811 (2025)
9. Chern, E., Su, J., Ma, Y., Liu, P.: Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. arXiv preprint arXiv:2407.06135 (2024)
10. Chow, W., Pan, J., Liang, Y., Zhou, M., Song, X., Jia, L., Zhang, S., Tang, S., Li, J., Zhang, F., et al.: Weave: Unleashing and benchmarking the in-context interleaved comprehension and generation. arXiv preprint arXiv:2511.11434 (2025)
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
12. Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., Wang, Y., Wang, C., Zhang, F., Zhao, Y., Pan, T., Li, X., Hao, Z., Ma, W., Chen, Z., Ao, Y., Huang, T., Wang, Z., Wang, X.: Emu3.5: Native Multimodal Models are World Learners. arXiv 2510.26583 (2025)
13. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Riquelme Ruiz, C., Minderer, M., Puigcerver, J., Evci, U., Kumar, M., Steenkiste, S.V., Elsayed, G.F., Mahendran, A., Yu, F., Oliver, A., Huot, F., Bastings, J., Collier, M., Gritsenko, A.A., Birodkar, V., Vasconcelos, C.N., Tay, Y., Mensink, T., Kolesnikov, A., Pavetic, F., Tran, D., Kipf, T., Lucic, M., Zhai, X., Keysers, D., Harmsen, J.J., Houlsby, N.: Scaling Vision Transformers to 22 Billion Parameters. In: ICML (2023)

14. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging Properties in Unified Multimodal Pretraining. arXiv 2505.14683 (2025)
15. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., Zhang, X., Ma, K., Yi, L.: DreamLLM: Synergistic Multimodal Comprehension and Creation. In: ICLR (2024)
16. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
17. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In: ICML (2024)
18. Ge, Y., Zhao, S., Zeng, Z., Ge, Y., Li, C., Wang, X., Shan, Y.: Making llama see and draw with seed tokenizer. arXiv preprint arXiv:2310.01218 (2023)
19. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. In: NeurIPS (2023)
20. He, X., Wei, L., Ouyang, J., Liao, M., Xie, L., Tian, Q.: EMMA: Efficient Multimodal Understanding, Generation, and Editing with a Unified Architecture. arXiv 2512.04810 (2025)
21. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment (2024)
22. Huang, R., Wang, C., Yang, J., Lu, G., Yuan, Y., Han, J., Hou, L., Zhang, W., Hong, L., Zhao, H., Xu, H.: ILLUME+: Illuminating Unified MLLM with Dual Visual Tokenization and Diffusion Refinement. arXiv 2504.01934 (2025)
23. Jiao, Y., Qiu, H., Jie, Z., Chen, S., Chen, J., Ma, L., Jiang, Y.G.: Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In: CVPR (2025)
24. Kou, S., Jin, J., Liu, Z., Liu, C., Ma, Y., Jia, J., Chen, Q., Jiang, P., Deng, Z.: Orthus: Autoregressive Interleaved Image-Text Generation with Modality-Specific Heads. In: ICML (2025)
25. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025)
26. Li, A., Wang, C., Fu, D., Yue, K., Cai, Z., Zhu, W.B., Liu, O., Guo, P., Neiswanger, W., Huang, F., et al.: Zebra-cot: A dataset for interleaved vision language reasoning. arXiv preprint arXiv:2507.16746 (2025)
27. Li, B., Wang, Z., Li, F., Xu, J., Guo, J., Pei, R., Li, X., Chen, Z.: Colorflux: A structure-color decoupling framework for old photo colorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2026)
28. Li, S., Gu, J., Liu, K., Lin, Z., Wei, Z., Grover, A., Kuen, J.: Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. arXiv preprint arXiv:2509.19244 (2025)
29. Li, Z., Li, H., Shi, Y., Farimani, A.B., Kluger, Y., Yang, L., Wang, P.: Dual diffusion for unified image generation and understanding. In: CVPR (2025)

30. Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., Zhao, W., Wu, J., Li, L., Tian, Z., Huang, W.: Mogao: An Omni Foundation Model for Interleaved Multi-Modal Generation. arXiv 2505.05472 (2025)
31. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. In: ICLR (2024)
32. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
33. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow Matching for Generative Modeling. In: ICLR (2023)
34. Liu, X., Wei, Y., Liu, M., Lin, X., Ren, P., Xie, X., Zuo, W.: Smartcontrol: Enhancing controlnet for handling rough visual conditions. In: European Conference on Computer Vision. pp. 1–17. Springer (2024)
35. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., yu, X., Zhao, L., Wang, Y., Liu, J., Ruan, C.: JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation. arXiv 2411.07975 (2025)
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (2024)
37. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: ICLR (2024)
38. Qu, L., Cheng, F., Yang, Z., Zhao, Q., Lin, S., Shi, Y., Li, Y., Wang, W., Chua, T.S., Jiang, L.: VINCIE: Unlocking in-context image editing from video. In: ICLR (2026)
39. Shazeer, N.: GLU Variants Improve Transformer. arXiv 2002.05202 (2020)
40. Shi, Q., Bai, J., Zhao, Z., Chai, W., Yu, K., Wu, J., Song, S., Tong, Y., Li, X., Li, X., et al.: Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. arXiv preprint arXiv:2505.23606 (2025)
41. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
42. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: RoFormer: Enhanced transformer with Rotary Position Embedding. Neurocomputing (2024)
43. Team, C.: Chameleon: Mixed-Modal Early-Fusion Foundation Models. arXiv 2405.09818 (2025)
44. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: A family of highly capable multimodal models. arxiv 2023. arXiv preprint arXiv:2312.11805 (2024)
45. Teed, Z., Deng, J.: RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In: ECCV (2020)
46. Tian, C., Zhu, X., Xiong, Y., Wang, W., Chen, Z., Wang, W., Chen, Y., Lu, L., Lu, T., Zhou, J., Li, H., Qiao, Y., Dai, J.: MM-Interleaved: Interleaved Image-Text Generative Modeling via Multi-modal Feature Synchronizer. arXiv (2024)
47. Wang, C., Chen, Z., Wei, Y., Jiang, T., Wu, X., Li, F., Zuo, W., Yao, H.: Creval: An automated interpretable evaluation for creative image manipulation under complex

instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2026)
48. Wang, X., Zhang, Z., Zhang, H., Lin, Z., Zhou, Y., Liu, Q., Zhang, S., Li, Y., Liu, S., Zheng, H., et al.: Hbridge: H-shape bridging of heterogeneous experts for unified multimodal understanding and generation. arXiv preprint arXiv:2511.20520 (2025)
49. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
50. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., Luo, P.: Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation. arXiv 2410.13848 (2024)
51. Wu, C., Lei, L., Li, F., Guo, C., Kong, D., Qin, X., Wang, Z., Cheng, M., Li, C.: Yose: You only select essential tokens for efficient dit-based video object removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2026)
52. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: OmniGen2: Exploration to Advanced Multimodal Generation. arXiv 2506.18871 (2025)
53. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
54. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564 (2025)
55. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809 (2025)
56. Yang, S., Ge, Y., Li, Y., Chen, Y., Ge, Y., Shan, Y., Chen, Y.C.: Seed-story: Multimodal long story generation with large language model. In: ICCV Workshops. pp. 1850–1860 (October 2025)
57. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment. In: CVPR (2022)
58. Zhang, B., Sennrich, R.: Root Mean Square Layer Normalization. In: NeurIPS (2019)
59. Zhang, C., Wang, J., Wang, Y., Liang, Y., Yang, X., Li, Z., Huang, H., Li, X.: Unimodel: A visual-only framework for unified multimodal understanding and generation. arXiv preprint arXiv:2511.16917 (2025)
60. Zhang, H., Qu, L., Liu, Y., Chen, H., Song, Y., Dong, Y., Sun, S., Li, X., Wang, X., Jiang, Y., Ye, H., Chen, B., Gao, Y., Liu, P., Liu, A., Yang, Z., Deng, Q., Xing, L., Liu, J., Wang, Z., Zhou, Y., Liu, M., Zhang, Y., He, Q., Hu, X., Qi, Z., Shao, J., Fu, Z., Wang, S., Chen, F., Chai, X., Wu, Z., Wang, Y., Yuan, Z., Du, D.K., Wu, X.: NextFlow: Unified Sequential Modeling Activates Multimodal Understanding and Generation. arXiv 2601.02204 (2026)
61. Zheng, K., He, X., Wang, X.E.: Minigpt-5: Interleaved vision-and-language generation via generative vokens. arXiv preprint arXiv:2310.02239 (2023)
62. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. In: ICLR (2025)