Wan-Weaver: Interleaved Multi-modal Generation via Decoupled Training

Jinbo Xing[1] Zeyinzi Jiang[1] Yuxiang Tuo[1] Chaojie Mao[1] Xiaotang Gai[1] Xi Chen[1]
Jingfeng Zhang[1] Yulin Pan[1] Zhen Han[1] Jie Xiao[1] Keyu Yan[1] Chenwei Xie[1]
Chongyang Zhong[1] Kai Zhu[1] Tong Shen[1] Lianghua Huang[1] Yu Liu[1] Yujiu Yang[2]
[1]Tongyi Lab [2]Tsinghua University

Figure 1. Showcase of the versatile abilities of *Wan-Weaver*, including interleaved text-image generation, reference-based image generation/editing, and text-to-image generation.

Abstract

Recent unified models have made unprecedented progress in both understanding and generation. However, while most of them accept multi-modal inputs, they typically produce only single-modality outputs. This challenge of producing interleaved content is mainly due to training data scarcity and the difficulty of modeling long-range cross-modal context. To address this issue, we decompose interleaved generation into textual planning and visual consistency modeling, and introduce a framework consisting of a planner and a visualizer. The planner produces dense textual descriptions for visual content, while the visualizer synthesizes images accordingly. Under this guidance, we construct large-scale textual-proxy interleaved data (where visual content is represented in text) to train the planner, and curate reference-guided image data to train the visualizer. These designs give rise to Wan-Weaver, which exhibits emergent interleaved generation ability with long-range textual coherence and visual consistency. Meanwhile, the integration of diverse understanding and generation data into planner training enables Wan-Weaver to achieve robust task reasoning and generation proficiency. To assess the model's capability in interleaved generation, we further construct a benchmark that spans a wide range of use cases across multiple dimensions. Extensive experiments demonstrate that, even without access to any real interleaved data, Wan-Weaver achieves superior performance over existing methods.

1. Introduction

Generative models are driving progress toward artificial general intelligence. Recent advances in large language models (LLMs) [39, 62], vision language models (VLMs) [3, 15], unified multi-modal models (UMMs) [10, 16], and image generation models [7, 22, 52] have enabled processing of multi-modal inputs, but usually with single-modal outputs. However, achieving human-like interaction requires the ability to generate multi-turn interleaved multi-modal outputs, which are crucial for applications in reasoning [6, 46], education [13, 30], and design [26, 54].

However, generating natural and reliable multi-modal content remains highly challenging, as the outputs across modalities must remain consistent. Early approaches [12, 28, 70, 81] fine-tune LLMs on image–caption datasets to acquire basic image generation capabilities; however, such training enables only elementary visual synthesis and fails to capture contextual dependencies. Recent unified multi-modal models (UMMs) [19, 57, 59] adopt an autoregressive next-token prediction paradigm or its combination with diffusion models, such as BAGEL [16] and Mogao [35], and are pre-trained on interleaved text–image data. However, the scarcity of large-scale, high-quality interleaved data leads to sparse supervision, making joint optimization unstable and hindering the model's ability to learn long-range contextual dependencies and modality transitions.

Given the difficulty of data curation, can we achieve interleaved generation without interleaved training data? We argue that the goal of **"interleaved coherence" could be decomposed into textual, visual, and cross-modal dimensions.** The textual coherence has already been satisfied by modern VLMs [3]. The visual coherence among images resembles that found in reference-based image generation/editing, for which abundant data exist or *can be synthesized*. Likewise, basic text–image alignment in cross-modal coherence can be effectively learned from large text-to-image corpora. However, **cross-modal coherence also requires planning capabilities**, *e.g.*, determining where an image should appear in the sequence and what it should portray to fit the long-range context and narrative flow.

In this way, the problem reduces to training a planner and learning each coherence dimension separately. We therefore propose *Wan-Weaver*, a MoT-architecture [16, 34] unified multi-modal model consisting of a planning expert and a visualization expert. This design enables decoupled training: The planner is initialized from a pre-trained VLM, and fine-tuned on large-scale textual-proxy planning and understanding data, where each image is located and represented by a *dense prompt*. The visualizer is then trained on reference-guided generation data, with the planner kept frozen to provide contextual features. During sequential inference, The planner processes the input together with the previously generated text–image context and produces visualization guidance and plain text, while the visualizer generates corresponding images conditioned on the guidance and visual references. Although the dense visual guidance produced by the planner provides useful high-level semantics for the current visualization step, the textual form of the dense prompt inevitably loses subtle contextual cues. Inspired by the notion of context windows [3], we introduce a *dense prompt context window* and further fine-tune the visualizer to improve the consistency of the generated content.

Moreover, our multi-task–trained planner exhibits emergent task reasoning across understanding and generation tasks, advancing toward more intelligent multi-modal generation. To support comprehensive evaluation, we further introduce WeaverBench, a benchmark specifically designed for interleaved generation and covering a broad range of everyday use cases. Extensive experiments demonstrate that our proposed method not only surpasses leading open-source counterparts but also achieves performance on par with the commercial model Nano Banana. Fig. 1 features our work. Our contributions are summarized as follows:

- We decompose the interleaved multi-modal generation problem and design a unified architecture comprising dedicated planning and visualization experts.
- We curate large-scale proxy cross-modal data to address the lack of interleaved supervision and develop a decoupled training strategy that substantially improves interleaved generation over contemporary baselines.
- We propose a benchmark for evaluating open-ended interleaved image-text generation, covering a wide range of daily use cases.

2. Related Work

Unified Multi-modal Models [10, 11, 45, 67, 75] aim to build a single architecture capable of both understanding and generation. They take images and text as input and produce either text or image as output. While current VLMs generally extend GPT-style LLMs [5, 50, 51] and adopt next-token prediction for multi-modal understanding [1, 3, 24, 31, 84], while state-of-the-art visual generators rely on diffusion modeling [17, 29]. This trend has spurred investigations into various paradigms for unifying multi-modal models, which can be roughly classified into three categories: autoregressive (AR) models, fused AR with diffusion models, and diffusion models. AR models [18, 28, 41, 49, 55–57, 67, 72, 77] follow the paradigm of next-token prediction in language models [62] by encoding images in continuous embeddings or discrete tokens. However, AR approaches still lag behind diffusion-based ones in visual fidelity [8, 32, 40, 60]. Fused AR with diffusion models integrates diffusion processes into language backbones, either using additional image decoder [10, 21, 25, 33, 36, 48, 61, 64, 65, 68, 69, 71, 76, 79] or employing a shared Transformer architecture [53, 73, 74, 80, 82].

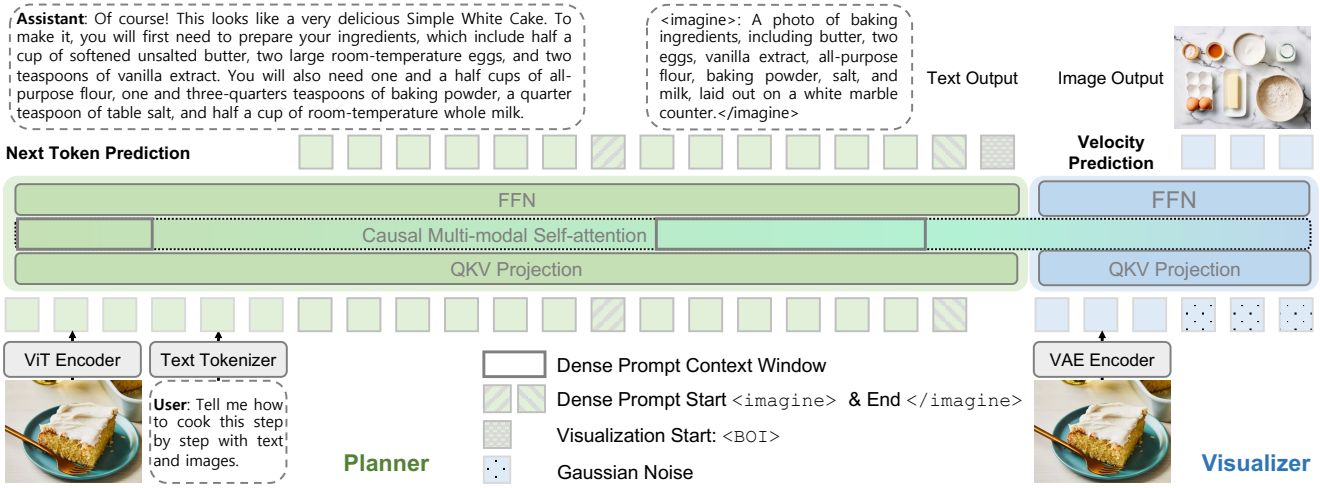

Figure 2. Overview of the inference process of *Wan-Weaver*. Given a prompt, the planner expert autoregressively generates plain text and dense prompts as visualization cues. Through causal multi-modal self-attention, the visualizer interacts with the planner, enabling it to synthesize images conditioned on the dense prompt context and visual references. The resulting text–image outputs are appended to the history and fed back into the planner, enabling an iterative interleaved generation process that maintains long-range contextual coherence.

Although this hybrid design improves visual quality, it often degrades multi-modal reasoning, as shared parameters must simultaneously optimize for both generation and understanding [34, 37]. Recent methods [16, 35] introduce MoT [34] architecture to successfully achieve better performance with separated Transformer parameters. Diffusion-based approaches [66, 75] attempt to completely abandon autoregressive architectures, pursuing unified vision-language modeling from the perspective of discrete diffusion or flow matching, achieving considerable results. Our unified model adopts the architecture of MoT and employs a decoupled training strategy for interleaved generation.

Interleaved Multi-modal Generation is a primary use case of unified models [11, 19, 27, 45, 76]. Pioneering AR frameworks [56, 77] demonstrated that models trained on interleaved sequences could generate both text and images, enabling diverse tasks such as text-to-image generation, visual understanding, and image editing. To enhance the visual realism, subsequent approaches [82] substituted image token-wise prediction with iterative denoising for superior image fidelity. However, these approaches are predominantly designed for single-turn generation, producing an isolated text or image output. Recent works [16, 28, 35] have attempted to pre-train models on collected interleaved data to enhance their multi-turn multi-modal generation capabilities. However, due to the difficulty in acquiring large-scale and reliable interleaved data, it is challenging for these models to learn complex long-range contextual dependencies. Consequently, this paradigm has yielded suboptimal results. In contrast, we explore a strategy of decoupled training within a unified model, leveraging large-scale synthetic data to achieve superior performance.

3. Method

3.1. Problem Definition

Interleaved multi-modal generation aims to produce a sequence of text and images from an input prompt. At the modal level, the distribution over this multimodal sequence can be written in a causal form:

$$\log P_\theta(\mathbf{x}) = \sum_{t=0}^{T} \log P_\theta(\mathbf{x}_{t+1} \mid \mathbf{x}_0, \ldots, \mathbf{x}_t),$$

where θ denotes the parameters of the model and $\mathbf{x}_t$ represents either a text or visual modality rather than an individual token. This formulation is general and encompasses a broad spectrum of existing tasks, including but not limited to text question answering, image captioning, visual question answering, text-to-image generation, and image editing, among others. While our model is also capable of handling these tasks (see Sec.4.4), they are not the primary focus of this work. Instead, we focus on interleaved text–image generation, with particular emphasis on the more common interleaved pattern, where textual and visual elements are produced alternately in a contextually coherent and semantically aligned manner.

3.2. Overview of Wan-Weaver

The overall architecture of Wan-Weaver is illustrated in Fig. 2. Motivated by our decomposition of interleaved generation into planning and visual coherence modeling, we adopt a unified mixture-of-transformers (MoT) [16, 34] framework comprising a planner and a visualizer expert, which work jointly to produce long-range coherent interleaved text–image content in a coordinated manner. The

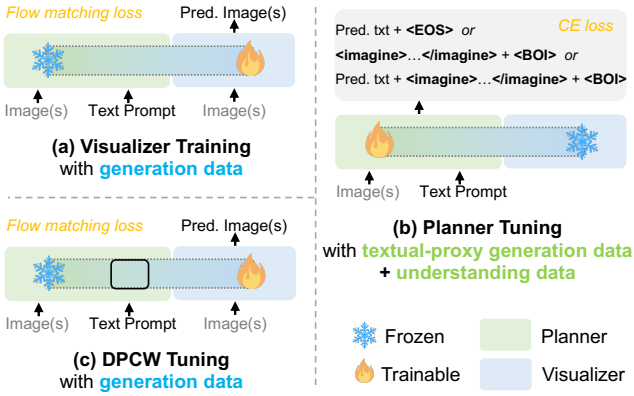

Figure 3. Illustration of our decoupled training strategy.

planner, instantiated as a VLM, handles modal reasoning and planning, deciding both the modality of the next output and the content of the image when one is required. Given a multi-modal user instruction, text is tokenized and images are encoded with a ViT encoder [3], and the resulting tokens are processed using generalized causal multi-modal self-attention to produce plain text responses. When deciding to transition to the image modality, the planner generates a dense prompt that provides rich visual guidance by aggregating the complex long-range multi-modal context.

The dense prompt triggers the visualizer, implemented as a Diffusion Transformer, which synthesizes the corresponding image. Inspired by context windows in language models [39], we introduce a Dense Prompt Context Window (DPCW) that extracts precise contextual features around the dense-prompt position. The visualizer then performs causal multi-modal self-attention over this window. This targeted conditioning leverages long-range context without introducing excessive noise, leading to better contextual grounding and cross-modal coherence, while VAE-encoded reference tokens help preserve fine details. Once generated, each text or image is appended to the history for conditioning, enabling Wan-Weaver to autoregressively produce interleaved sequences from any mixture of modalities.

3.3. Decoupled Training Strategy

Interleaved generation is inherently challenging, as it requires complex cross-modal reasoning over long contexts and demands outputs that are logically consistent, semantically coherent, and aligned in fine-grained details among images. A straightforward approach is to pre-train a unified model on large-scale interleaved data; however, such high-quality data are scarce, making the model fail to learn reliable long-range contextual dependencies and modality transitions, and produce misaligned or logically inconsistent cross-modal outputs. To address this, we decompose interleaved generation into planning and visual coherence and implement a planner–visualizer architecture that supports decoupled training: the planner learns contextual planning from large-scale synthetic textual-proxy and understanding data, while the visualizer learns visual coherence from abundant reference-guided generation data. To leverage strong open-source VLMs, we initialize the planner with the pre-trained QWen2.5-VL [3]. As no publicly available diffusion transformer matches our visualizer's architecture, the visualizer is trained from scratch.

Visualizer Training. The visualizer is designed to synthesize images aligned with the planner's visual guidance and to preserve strong reference-driven consistency across images through causal attention. Achieving this requires substantial generative data. To satisfy the coherence demands of sequential interleaved generation, we decompose the problem into distinct forms of guidance coherence spanning text, single-image, and multi-image contexts.

For text-guidance coherence, the visualizer must align with the semantic space of the planner. We therefore collect large-scale text–image pairs, including both simple and reasoning-heavy descriptions/instructions, to ensure broad coverage. After the first image is generated, the model must also reference previously generated text and images to maintain contextual consistency. To this end, we prepare extensive single-image reference data. Similarly, multi-image reference data are incorporated to further enhance long-range consistency across sequences of images. Using these coherence-oriented datasets, we freeze the planner and independently train the visualizer with flow-matching loss [38] to achieve multi-modal consistency under diverse conditions, as shown in Fig. 3 (a).

Planner Tuning. The planner expert is responsible for digesting complex multi-modal inputs and producing text responses. Crucially, it must infer when to produce text vs. an image and what specific visual content should be generated given the surrounding context—capabilities that our initialized planner only partially supports. Fine-tuning is therefore required, but this is challenging due to the scarcity of large-scale, high-quality interleaved data.

To address this issue, we synthesize large-scale high-fidelity *textual-proxy interleaved data* using top-tier LLMs and VLMs. Each image placeholder is tagged with `<BOI>` and accompanied by a detailed caption describing the intended visualization. These fine-grained annotations, called *dense prompts* and enclosed in `<imagine></imagine>`, provide richer image-specific guidance than the 'sparse' surrounding context and align with the visualizer's textual conditioning during training.

Training with such data offers several advantages: (1) it leverages the inherent ability of large VLMs to integrate long-range multi-modal information, enabling precise image-generation guidance from purely textual dense prompts and effectively benefiting generation with understanding; (2) it exploits the semantic equivalence of images

and text in the language modeling space, allowing LLMs to learn when to transition modalities without relying on real interleaved data, thus alleviating data scarcity; and (3) it leads to more stable optimization, as joint-training with denoising generation objectives typically introduces gradient interference [48]. Moreover, this textual-proxy mechanism can generalize to other generation tasks, forming textual-proxy generation data. To further equip the planner with automatic task reasoning (*e.g.*, understanding, text-to-image generation, image editing, interleaved generation), we jointly train it on these proxy datasets together with conventional understanding data, as shown in Fig. 3 (b).

Once trained, our unified model differs from prior UMMs [16, 21], which depend on explicit task-specific system prompts or rigid if-else logic to predefine the output modality by users. In contrast, our approach allows the model to implicitly infer task intent from user prompt.

Dense Prompt Context Window Tuning. While the generated dense prompt provides a rich textual description of the intended image, it remains a purely text representation and inevitably suffers from information loss, failing to capture subtle contextual nuances. Thus, we introduce the Dense Prompt Context Window (DPCW), a mechanism designed to enhance contextual grounding for image generation based on the generated dense prompt. Specifically, during the visualization process, a self-attention window is defined on top of the original causal self-attention centered around the position where the dense prompt is produced. Only the contextual information within this window participates in self-attention interactions with the visualizer. Moreover, since the ViT features of the input image are crucial for maintaining semantic faithfulness during generation [16], this visual context is also incorporated.

In this way, DPCW not only encapsulates the semantic richness of the dense prompt but also aggregates the *preceding contextual information* accumulated through layer-wise causal self-attention. This results in a more comprehensive context that is continuously propagated across the sequential planning–visualization process, thereby improving the coherence and consistency of the generated content. As illustrated in Fig. 3 (c), we perform an additional DPCW tuning stage to adapt the model to this context-window-aware conditioning. In this stage, only the visualizer is fine-tuned.

3.4. Data Curation

To support the decoupled training strategy, we curate and synthesize large-scale datasets tailored to the objectives of each training stage.

Visualizer tuning data. We curated a diverse corpus combining public text-to-image and image editing datasets. The text instructions were augmented through rewriting to simulate various user input scenarios. Beyond standard text–image pairs and public image-to-image datasets, we further construct large-scale image-reference-guided generation data from two sources for visual coherence learning. First, we extract high-quality key frames from videos and use a VLM to generate detailed textual descriptions as well as instructions describing the changes across selected frame combinations. Because video frames typically exhibit smooth temporal continuity, most variations are localized rather than drastic. As a result, this source mainly provides data for learning fine-grained local edits and high-coherence reference-based generation. Second, to obtain more general reference-based generation capability that can match that with real interleaved data, we cluster homologous images in our database using SigLIP [78] and apply a VLM to create structured descriptions spanning both single-image and multi-image reference scenarios. Unlike video frames, these clusters exhibit far richer diversity, including reference-based learning on style, material, pose, expression, and clothing, which provides highly versatile reference supervision. Together, these datasets enhance the visualizer's ability to produce consistent and semantically coherent outputs under diverse reference conditions.

Planner tuning data. To preserve the language modeling capacity of the underlying VLM, we collect high-quality text-only and image–text understanding data. To endow the planner with interleaved generation planning capability, we constructed *textual-proxy data*, where visual content was replaced by detailed captions annotated with `<imagine>`. Specifically, three types of data were synthesized: (1) large-scale user query (text only)–interleaved article pairs generated by prompting the top-tier LLMs with category or tag keywords; (2) user query (with images)–interleaved article pairs built around arbitrary images from the database by VLMs; and (3) multi-image data where each image was first captioned by a VLM, then organized into coherent interleaved narratives, and finally refined for logical and stylistic consistency. This refinement step is essential, as independently generated VLM captions may introduce inconsistencies due to inherent sampling variability. Note that the dense prompts in the textual-proxy data exhibit substantial diversity: they range from short phrases to long, detailed descriptions, and may also specify changes relative to particular reference images. The same proxy strategy was also applied to non-interleaved generation tasks such as text-to-image and reference-based image generation, improving the reasoning and user intent comprehension ability.

4. Experiment

4.1. Implementation Details

We initialize the planner with an in-house Qwen2.5-VL-32B-Think [3] model and implement a twin-structured, DiT-based visualizer, trained from scratch. A frozen visual encoder—VAE from Wan2.2 [63] plus the Qwen2.5-

Table 1. Quantitative comparison with existing state-of-the-art methods on interleaved generation benchmark OpenING [83].

Method	Completeness	Quality	Richness	Correctness	Human Alignment	IT Coherency	Multi-step Consistency	Overall
NExT-GPT	3.89	4.25	3.35	3.61	5.35	3.32	3.85	3.95
MiniGPT-5	3.91	4.50	3.61	3.63	5.51	3.56	4.10	4.12
Orthus	4.43	4.30	3.71	4.15	4.80	3.51	4.20	4.16
Show-o	4.37	4.79	3.83	3.76	5.78	4.04	4.33	4.41
VILA-U	5.60	5.14	4.68	4.78	5.69	4.74	4.79	5.06
SEED-LLaMA	5.59	5.50	4.61	4.59	6.50	4.43	5.13	5.19
Anole	6.27	6.02	5.28	5.06	6.91	4.90	5.81	5.75
Emu3	5.90	5.96	5.52	5.43	6.47	5.66	5.37	5.76
SEED-X	5.65	6.07	4.92	5.77	7.03	5.72	5.72	5.84
Gemini+Flux	7.58	7.26	6.48	7.03	7.98	6.98	7.33	7.23
GPT-4o+DALL-E3	8.66	8.01	7.42	7.98	8.77	8.15	8.38	8.20
Nano Banana	9.34	**8.58**	8.00	**9.17**	**8.88**	**9.27**	**8.70**	**8.85**
Wan-Weaver (Ours)	**9.41**	8.32	**8.03**	8.90	8.69	8.78	8.56	8.67

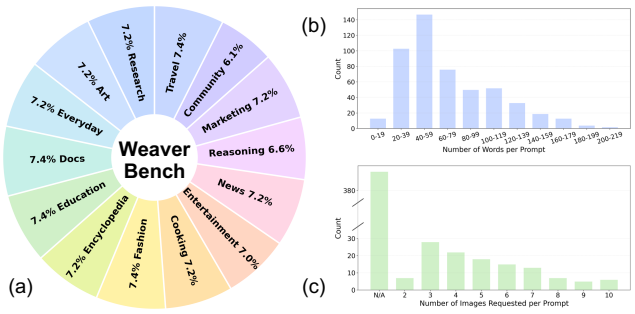

Figure 4. Statistics of WeaverBench. (a) Topic distribution across 14 everyday categories. (b) Prompt length distribution. (c) Distribution of the number of images requested per prompt.

VL ViT—is used throughout all experiments. The visualizer is trained for 9.6T tokens with AdamW optimizer, using a learning rate that decays from 5×10^{-5} to 2.5×10^{-5}; images keep native aspect ratios with resolutions spanning $\sim 196^2$ to 1440^2, and the share of high-resolution data is progressively increased. Training alternates over three stages: text-to-image, text-image-to-image, and text-multi-image-to-image. The planner is tuned over 35.72G tokens with AdamW at 7×10^{-6}, using a 5:1 generation-to-understanding sampling ratio. DPCW is realized via an attention-masking strategy, and 3D RoPE [63] is employed.

4.2. Benchmark: WeaverBench

Prior benchmarks [2, 42] target only a narrow set of topics. While recent efforts aim for broader coverage, some of their user prompts nevertheless involve non-interleaved tasks (*e.g.*, pure understanding or single-image editing) [83] or remain overly templated [9]. We therefore introduce a compact, interleaved-generation–focused benchmark, characterized by flexible user queries and diverse generation scenarios grounded in everyday use, aiming to foster more rigorous and practically meaningful assessment.

With the assistance of multiple AI agents, we collaboratively brainstormed and identified a comprehensive set of everyday scenarios that demand interleaved image–text generation. These insights were consolidated into 14 primary categories. To construct corresponding test prompts, we manually collected and crafted user queries from diverse sources (*e.g.*, social media, search engines, and public knowledge bases) to ensure broad topical coverage. The resulting prompts exhibit varying levels of specificity, ranging from short queries with only a few words to complex requests that specify the number, order, and semantic content of each image to be generated. In total, the benchmark comprises 512 test cases, evenly split between image-conditioned and text-only prompts. Fig. 4 presents the statistics of our WeaverBench.

We carefully identify the primary aspects of interleaved generation and design corresponding evaluation metrics, including Prompt Adherence, Narrative Coordination, Content Consistency, and Completeness. As demonstrated in [83], GPT-4o [47] aligns well with human preferences. Following this, we develop finer-grained scoring rules to obtain more reliable assessments.

Table 2. Quantitative comparison on WeaverBench. PA: Prompt Adherence, NC: Narrative Coordination, CC: Content Consistency, IC: Image Consistency, CP: Completeness.

Method	PA	NC	CC	IC	CP	Overall
Orthus	2.47	1.88	1.69	1.51	1.91	1.89
Anole	4.14	3.76	3.77	3.42	3.64	3.74
Emu3.5	7.65	7.55	7.56	7.50	7.41	7.53
Nano Banana	8.53	8.19	**8.53**	**8.38**	8.29	8.38
Wan-Weaver (Ours)	**8.71**	**8.33**	8.50	8.13	**8.46**	**8.43**

4.3. Interleaved Image-Text Generation

To evaluate the quality of interleaved generation, we employ the OpenING benchmark [83], a dataset comprising over 2,000 samples, along with the seven metrics proposed in the original paper. We evaluate our method against a broad range of existing approaches, which can be categorized into four groups. (1) Integrated pipelines combine indepen-

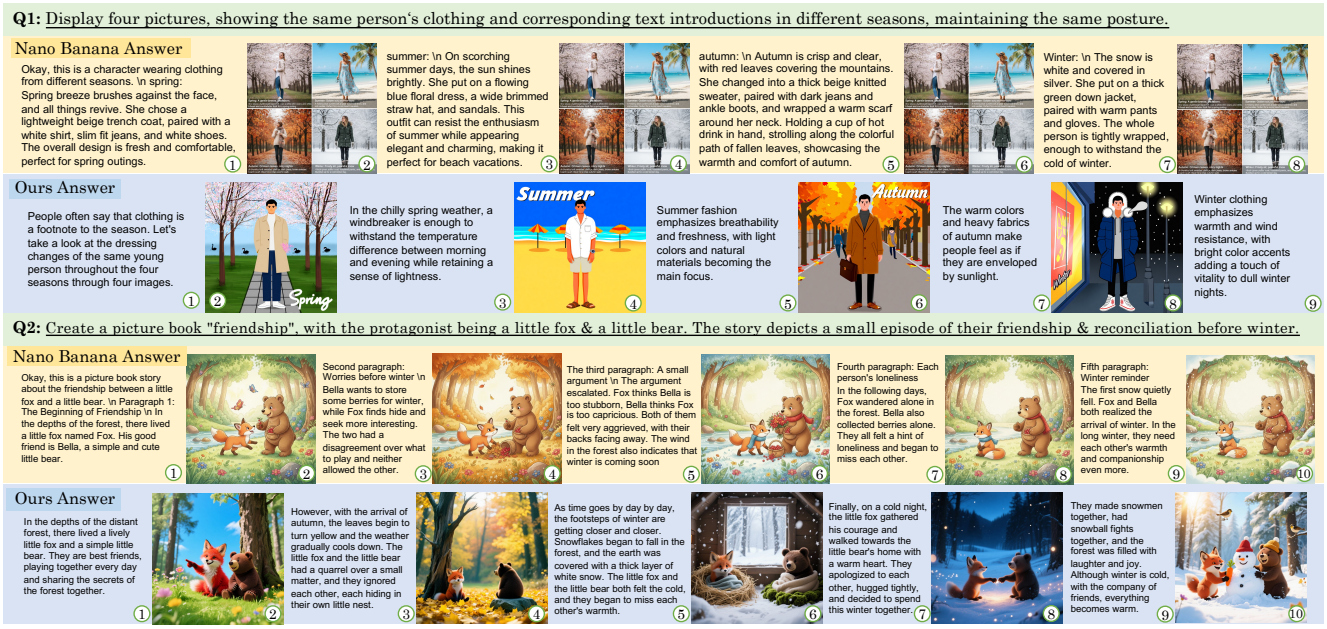

Figure 5. Qualitative comparison with the state-of-the-art commercial system Nano Banana on interleaved text–image generation.

dent text and image generation models: GPT-4o+DALL·E-3 [4, 47] and Gemini+Flux [29, 58]. (2) Two-stage generators that adopt a unified model architecture but generate textual and visual content sequentially in two distinct stages: Emu3 [67], SEED-X [20], VILA-U [72], and Show-o [73]. (3) End-to-end generators produce interleaved image–text content within a single autoregressive process: Orthus [28], NExT-GPT [70], MiniGPT-5 [81], SEED-LLaMA [19], and Anole [12]. (4) Commercial interleaved generators that directly support text–image interleaving: Gemini-2.5-Image (Nano Banana) [23]. Table 1 shows that our method clearly surpasses all open-source and integrated-pipeline baselines. While the top-tier commercial model Nano Banana—whose implementation details are not publicly available—retains a small overall lead, our method is highly competitive and outperforms it on several metrics. We further note that Nano Banana's strong image–text coherence and multi-step consistency may arise from repeated or highly similar image outputs, as illustrated in Fig. 5. In addition, we further assess representative unified models (with general interleaved capabilities [12, 14, 23, 28]) on our WeaverBench to examine performance across diverse daily use cases. As shown in Table 2, the results reveal a similar conclusion, further demonstrating the superior performance of our method.

4.4. Single Modality Generation

Our model is a unified multi-modal generation framework trained on diverse understanding and generation data, enabling it to support both interleaved multi-modal generation and standard single-modality tasks. For comparison, we include understanding-only models (*i.e.*, InternVL3-38B [84], Ovis2-34B [44], and Qwen2.5-VL-32B [3]), generation-only models (*i.e.*, FLUX.1-dev [29] and Step1X-Edit [43]), and unified models (*i.e.*, BAGEL [16] and UniWorld-V1 [36]). As shown in Table 3, it delivers strong performance across understanding, image generation, and editing benchmarks, markedly outperforming previous unified and specialized generation models. The qualitative results are shown in Fig. 1.

Table 3. Comparison across single-modality generation tasks (understanding, image generation, and editing). †: Our in-house base model with thinking mode (enabled only for understanding).

Model	Understanding		Image Generation		Image Editing	
	MMMU	MathVista	GenEval	DPG	ImgEdit	GEdit–EN
InternVL3-38B	69.7	76.3	–	–	–	–
Ovis2-34B	66.7	76.1	–	–	–	–
Qwen2.5-VL-32B†	**75.1**	**84.7**	–	–	–	–
FLUX.1-dev	–	–	0.66	84.0	–	–
Step1X-Edit	–	–	–	–	3.06	6.70
Unified Models						
Bagel	55.3	73.1	0.88	85.07	3.20	6.52
UniWorld-V1	58.6	–	0.84	81.38	3.26	4.85
Wan-Weaver (Ours)	74.9	84.3	**0.89**	**87.21**	**4.31**	**7.39**

4.5. Ablation Studies

We perform the ablation studies using the 7B variant owing to computational limitations.

Decoupled Training. To study the superiority of our decoupled training strategy, we compare it with the naive joint-training strategy on the full datasets. As shown in Fig. 6, where P+V indicates joint-training of the planner and visualizer, and V (T2I+SI2I+MI2I) denotes visualizer tuning, the decoupled training setting steadily reduces the vision

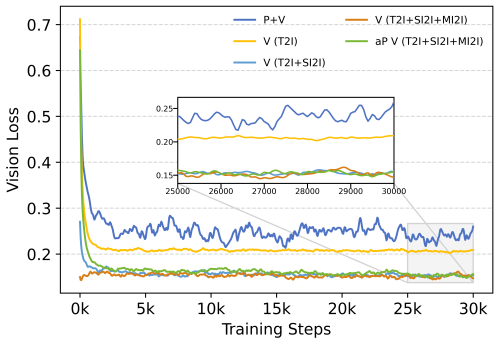

Figure 6. Loss curves of different training strategies.

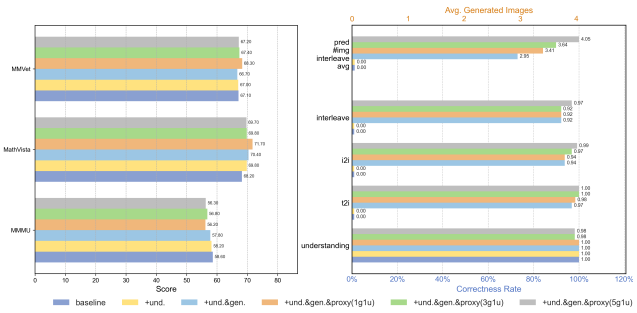

Figure 7. Impact of training on different types of data on our planner. (left) Performance on understanding metrics. (right) Token prediction statistics per task.

loss from around 0.25 to 0.15 and yields a much smoother optimization trajectory. This behavior is expected: when the planner and visualizer are optimized jointly, the non-negligible semantic and distribution gap between textual and visual modalities introduces instability, leading to misaligned text–image generation and degraded image quality. In contrast, isolating visualizer training avoids this cross-task interference, resulting in more stable convergence.

Feature Modeling in Planner. To investigate how planning levels affect understanding capability, we perform planner tuning on various data compositions. The baseline is a VLM-initialized planner without planning ability. As shown in Fig. 7, we compare three variants: (1) **+und.&gen.&proxy** provides comprehensive planning including dense prompts; (2) **+und.&gen.** preserves basic understanding with generative planning; and (3) **+und.** maintains only understanding data.

Fig. 7 (left) shows that understanding performance remains stable across configurations, indicating that planning does not compromise core understanding competency. We further evaluate modality-specific patterns via structural plan token accuracy. The planner must correctly emit <BOI> tokens: none for understanding, exactly one for T2I/I2I, and at least one for interleaved generation. As shown in Fig. 7 (right), without generation data, the model fails to emit image signals. However, as the generation-oriented data ratio increases (from 1g1u to 5g1u), planning

Figure 8. Visual comparison of the results generated by different variants of our method.

proficiency improves substantially, with a gradual increase in image starting tokens for interleaved tasks. Balancing planning reliability and understanding stability, we adopt the 5g1u ratio as the final composition for planner tuning.

Coherence Modeling in Visualizer. We constructed several training variants of our visualizer to investigate the effectiveness of our coherence modeling. Specifically, we train the visualizer using three settings: (1) only text–image paired data (T2I data), (2) T2I combined with single-image-to-image data (T2I+SI2I data), and (3) multi-image-to-image data added on top of all previous data (T2I+SI2I+MI2I data), which is our strategy. The corresponding qualitative result is shown in Fig. 8. We can see that T2I data-only training provides basic text–image alignment but lacks any reference ability, failing to generate the second image. Adding single-image reference data improves appearance preservation across steps, while introducing multi-image reference data further strengthens long-range visual coherence, enabling consistent object identity, style, and detail across multiple generated images.

As shown in Fig. 8 (right), the model without the dense prompt tends to generate repetitive and contextually irrelevant images. In contrast, using the dense prompt results in significantly more diverse outputs. The integration of our DPCW further enhances the model's ability to adhere to user instructions, particularly in long-context scenarios.

5. Conclusion

We presented Wan-Weaver, a unified multi-modal model with planner-visualizer architecture for interleaved text–image generation under limited interleaved supervision. By decomposing interleaved coherence into planning and visual coherence, synthesizing large-scale textual-proxy interleave data, and adopting a decoupled learning strategy, it effectively learns long-range contextual dependencies and generates consistent multi-modal content. Extensive experiments demonstrate our model substantially outperforms existing open-source approaches and delivers performance competitive with top-tier commercial demos.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 2

[2] Jie An, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo Luo. Openleaf: A novel benchmark for open-domain interleaved image-text generation. In *ACM MM*, 2024. 6

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 4, 5, 7

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*, 2(3):8, 2023. 7

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 2

[6] Ju-Seung Byun, Jiyun Chun, Jihyung Kil, and Andrew Perrault. Ares: Alternating reinforcement learning and supervised fine-tuning for enhanced multi-modal chain-of-thought reasoning through diverse ai feedback. *arXiv preprint arXiv:2407.00087*, 2024. 2

[7] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. 2

[8] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 2

[9] Dongping Chen, Ruoxi Chen, Shu Pu, Zhaoyi Liu, Yanru Wu, Caixi Chen, Benlin Liu, Yue Huang, Yao Wan, Pan Zhou, et al. Interleaved scene graphs for interleaved text-and-image generation assessment. *arXiv preprint arXiv:2411.17188*, 2024. 6

[10] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 2

[11] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 2, 3

[12] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024. 2, 7

[13] Daniel Claman, Emre Sezgin, et al. Artificial intelligence in dental education: opportunities and challenges of large language models and multimodal foundation models. *JMIR medical education*, 10(1):e52346, 2024. 2

[14] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025. 7

[15] Google Deepmind. Gemini2.5. *Gemini2.5*. https://gemini.google.com/, 2025. 2

[16] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2, 3, 5, 7

[17] 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 *ICML*, 2024. 2

[18] Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025. 2

[19] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023. 2, 3, 7

[20] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 7

[21] Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025. 2, 5

[22] Google. Imagen4. *Imagen4*. https://deepmind.google/models/imagen/, 2025. 2

[23] Google. Gemini-2.5-image. https://gemini.google/overview/image-generation/, 2025. 7

[24] Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Tai-Wei Chang, Jingdong Chen, et al. M2-omni: Advancing omni-mllm for comprehensive modality support with competitive performance. *arXiv preprint arXiv:2502.18778*, 2025. 2

[25] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv preprint arXiv:2504.01934*, 2025. 2

[26] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, pages 919–933, 2023. 2

[27] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. In *NeurIPS*, 2023. 3

[28] Siqi Kou, Jiachun Jin, Zhihong Liu, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. *arXiv preprint arXiv:2412.00127*, 2024. 2, 3, 7

[29] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2, 7

[30] Ehsan Latif, Gengchen Mai, Matthew Nyaaba, Xuansheng Wu, Ninghao Liu, Guoyu Lu, Sheng Li, Tianming Liu, and Xiaoming Zhai. Artificial general intelligence (agi) for education. *arXiv preprint arXiv:2304.12479*, 1:1–34, 2023. 2

[31] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[32] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, 2024. 2

[33] Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, and Li Yuan. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025. 2

[34] Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024. 2, 3

[35] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025. 2, 3

[36] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 2, 7

[37] Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*, 2024. 3

[38] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 4

[39] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 2, 4

[40] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, 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. 2

[41] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 2

[42] Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. *arXiv preprint arXiv:2406.14643*, 2024. 6

[43] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 7

[44] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025. 7

[45] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *CVPR*, 2025. 2, 3

[46] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. In *NeurIPS*, 2023. 2

[47] OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/, 2025. Accessed: 2025-10-11. 6, 7

[48] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 2, 5

[49] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *CVPR*, 2025. 2

[50] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI CDN*, 2018. 2

[51] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[52] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 2

[53] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024. 2

[54] Francesco Stella, Cosimo Della Santina, and Josie Hughes. How can llms transform the robotic design process? *Nature machine intelligence*, 5(6):561–564, 2023. 2

[55] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 2

[56] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024. 3

[57] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2

[58] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7

[59] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024. 2

[60] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*, 2024. 2

[61] Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *ICCV*, 2025. 2

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[63] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 5, 6

[64] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. In *ICCV*, 2025. 2

[65] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025. 2

[66] Jin Wang, Yao Lai, Aoxue Li, Shifeng Zhang, Jiacheng Sun, Ning Kang, Chengyue Wu, Zhenguo Li, and Ping Luo. Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. *arXiv preprint arXiv:2505.20147*, 2025. 3

[67] 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. 2, 7

[68] Hongyang Wei, Baixin Xu, Hongbo Liu, Cyrus Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, et al. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model. *arXiv preprint arXiv:2509.04548*, 2025. 2

[69] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 2

[70] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *ICML*, 2024. 2, 7

[71] Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025. 2

[72] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 2, 7

[73] 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. 2, 7

[74] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 2

[75] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025. 2, 3

[76] Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, et al. X-vila: Cross-modality alignment for large language model. *arXiv preprint arXiv:2405.19335*, 2024. 2, 3

[77] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multimodal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023. 2, 3

[78] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 5

[79] Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*, 2025. 2

[80] Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024. 2

[81] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023. 2, 7

[82] Chunting Zhou, LILI YU, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025. 2, 3

[83] Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *CVPR*, 2025. 6

[84] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2, 7