Loom: Diffusion-Transformer for Interleaved Generation

Mingcheng Ye[1] Jiaming Liu[2] Yiren Song[3†]
[1]Beijing Institute of Technology [2]Independent Researcher [3]National University of Singapore

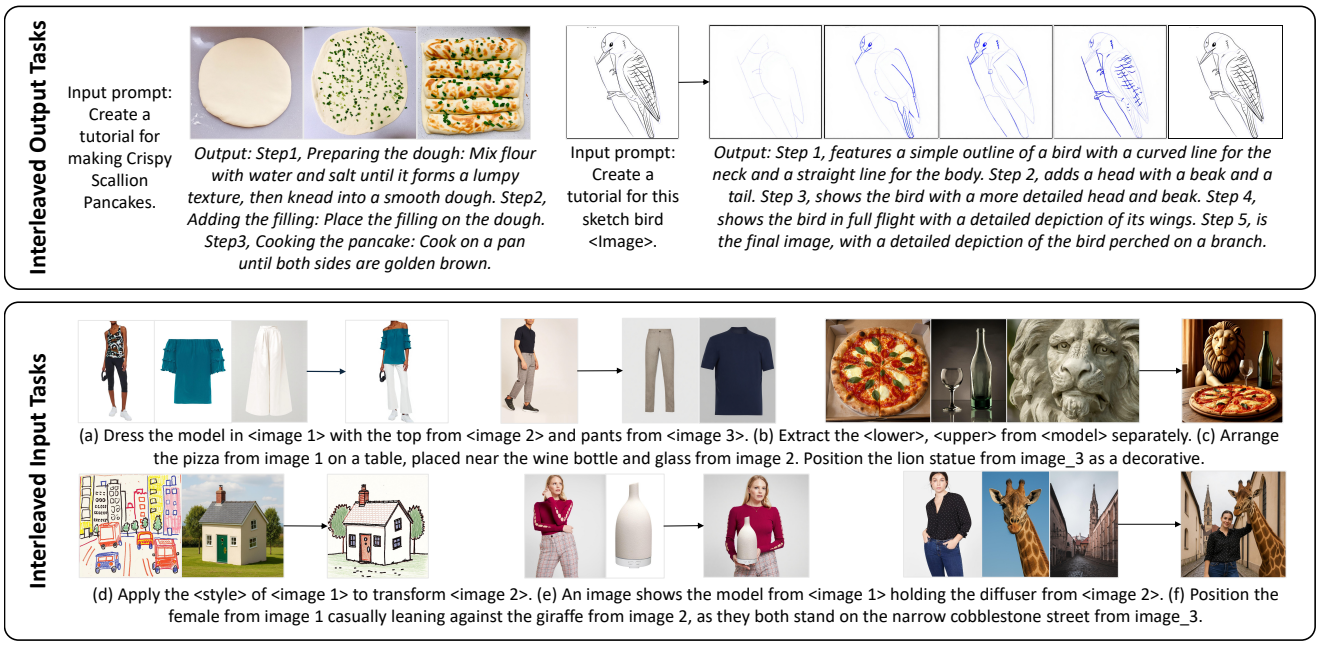

Figure 1. Showcases of Loom's interleaved text-image generation. Interleaved input tasks involve composing reference images into one scene or style transfer. Interleaved output tasks produce text–image sequences from a prompt, including cooking tutorials or drawing guide.

Abstract

Interleaved text–image generation aims to jointly produce coherent visual frames and aligned textual descriptions within a single sequence, enabling tasks such as style transfer, compositional synthesis, and procedural tutorials. We present Loom, a unified diffusion–transformer framework for interleaved text–image generation. Loom extends the Bagel unified model via full-parameter fine-tuning and an interleaved architecture that alternates textual and visual embeddings for multi-condition reasoning and sequential planning. A language-planning strategy first decomposes a user instruction into stepwise prompts and frame embeddings, which guide temporally consistent synthesis. For each frame, Loom conditions on a small set of sampled prior frames together with the global textual context, rather than concatenating all history, yielding controllable and efficient long-horizon generation. Across style transfer, compositional generation, and tutorial-like procedures, Loom delivers superior compositionality, temporal coherence, and text–image alignment. Experiments demonstrate that Loom substantially outperforms the open-source baseline Anole, achieving an average gain of 2.7 points (on a 5-point scale) across temporal and semantic metrics in text-to-interleaved tasks. We also curate a 50K interleaved tutorial dataset and demonstrate strong improvements over unified and diffusion editing baselines. Project page: https://github.com/Plantian/Loom.

1. Introduction

Open-source unified generative models like Bagel [12], Show-o [58], Janus-Flow [34], BLIP3-o [4], UniWorld [26] have shown that diverse visual tasks, spanning image editing, stylization, and layout-aware synthesis, can be handled within a single diffusion transformer. Yet most unified systems remain confined to single-turn, single-modality inputs: they either render an image from text or edit one reference in isolation.

† Corresponding author.

This fundamental limitation, however, extends far beyond simple generation as shown in Fig 1. A vast and challenging class of real-world scenarios demands reasoning over interleaved, mixed-modality sequences. These N-to-M tasks, which require models to consume and produce multiple, related inputs and outputs, include: (1) Procedural Generation: Producing step-by-step tutorials where visual frames and textual explanations are interleaved to guide a user, such as in cooking guides or artistic workflows. (2) Compositional Reasoning: Synthesizing a single, coherent scene from multiple, disparate visual and textual inputs, or the inverse, decomposing a scene into its constituent parts, for applications like virtual try-on. (3) Multi-Reference Generation: Transforming a content image based on the semantic or stylistic properties of several reference images, such as in complex style transfer. Current open-source frameworks lack a unified mechanism to handle this full spectrum of multi-modal, multi-turn reasoning.

In contrast, proprietary multimodals such as GPT-4o [38], Doubao [13], and Gemini [10] have demonstrated strong proficiency in such interleaved, multi-turn scenarios. These closed-source systems adaptively handle mixed-modality inputs and outputs across conversational contexts, maintaining semantic coherence over extended interactions. However, their proprietary nature restricts research transparency and limits customization for domain-specific applications.

To address this comprehensive gap, we introduce Loom, a unified model that treats text and images as sequentially composable elements within one transformer, enabling interleaved inputs and outputs. Loom is designed to target this full class of N-to-M (N inputs to M outputs) interleaved multi-modal generation tasks. It supports (1) **Procedural and tutorial generation**, producing step-by-step visual tutorials with aligned text. It handles (2) **Compositional generation and decomposition**, combining elements for virtual try-on or breaking images into their parts. Finally, it performs (3) **Style transfer** by conditioning on content images, style references, and text instructions.

To realize this unified approach, Loom treats text and image embeddings as sequentially composable elements within a shared latent space. We introduce a dual set of conditioning mechanisms to manage the complexity of N-to-M tasks. For procedural tasks, a language-planning strategy decomposes global instructions into local steps, which are associated with temporal frame embeddings and sparse historical frame sampling to maintain long-horizon coherence. For compositional and stylistic tasks, control is achieved via learnable entity tokens for structured grounding.

We train Loom by full-parameter fine-tuning on the Bagel backbone, effectively expanding its native capability from single-turn synthesis to complex, multi-turn interleaved generation. A key contribution is our curation of a new 50K interleaved tutorial dataset spanning drawing, cooking, and assembly workflows. Extensive evaluations show that Loom attains high-fidelity interleaved generation with accurate text-image alignment, strong temporal consistency, and robust multi-condition control. In short, our contributions are as follows:

(1) We propose Loom, a unified diffusion-transformer framework for interleaved text–image generation, supporting style transfer, compositional synthesis, and procedural tutorials within a single model.

(2) We introduce a unified control and conditioning mechanism for N-to-M tasks, including a language-planning strategy and sparse historical frame sampling for temporal coherence, and learnable entity tokens for structured compositional grounding.

(3) We curate a 50K interleaved tutorial dataset and present comprehensive experiments demonstrating Loom's superior compositionality, temporal coherence, and text–image alignment. By using this dataset, Loom substantially outperforms the open-source baseline Anole, achieving an average gain of 2.7 points (on a 5-point scale) across temporal and semantic metrics in text-to-interleaved tasks.

2. Related Work

2.1. Unified Models

Recent progress toward text–image unified modeling [54, 64] aims to bridge understanding and generation within a single transformer. Existing frameworks can be broadly categorized by backbone design into diffusion-based models, autoregressive multimodal large language models (MLLM-AR), and MLLM (AR+diffusion) architectures. Diffusion models [65] excel at pixel-level synthesis but often lack bidirectional reasoning for cross-modal understanding. Pure autoregressive MLLMs [46, 51, 68, 71] enable tighter integration between text and image tokens via next-token prediction, yet typically trade off visual fidelity. MLLM (AR+diffusion) frameworks [58] inject diffusion decoders into transformer backbones, improving generation while facing potential cross-task interference between modalities.

Our work builds on this unified modeling trend using **Bagel** [12], an MLLM (AR+diffusion) decoder-only transformer with Mixture-of-Experts (MoE) layers for semantic reasoning and visual synthesis. Bagel adopts a Vision Transformer (ViT) based hybrid visual encoder [49] to encode images into visual tokens for multimodal understanding, and a Variational Autoencoder (VAE) with rectified flow [21, 28] for high-fidelity image generation within a shared attention space, providing a strong foundation for the interleaved multimodal tasks studied in this work.

2.2. Procedural and Tutorial Generation

Procedural or tutorial generation spans storytelling, creative workflows, and instructional content, but most re-

mains largely text-centric [19, 24, 42] without unified bidirectional text–image modeling. Conventional text/image-to-painting pipelines [3, 43, 44, 47] often produce coarse steps lacking aligned textual logic, and recipe datasets like RecipeGen [62] yield unimodal outputs. In the realm of unified models, Anole [7] represents a state-of-the-art open-source baseline but is constrained to text-conditioned inputs. MM-Interleaved [48] adds visual tokens but struggles with error accumulation in long horizons. While concurrent approaches like OneFlow [37] and Orthus [20] improve speed, they typically compromise fine-grained stepwise control.

In contrast, Loom handles both text- and image-conditioned interleaved generation with stable progression, balanced alignment, and precise procedural adherence enabled by our planning-first strategy.

2.3. Multi-Reference Images Generation

Interleaved generation necessitates reasoning over multiple visual references simultaneously. One direction is **style transfer**, where a style reference and a content image yield stylized yet structure-preserving outputs. While diffusion-based approaches [16, 30–33, 35, 36, 45] effectively adapt diffusion backbones (e.g., FLUX [21]) via Low-Rank Adaptation (LoRA) [18], they rely heavily on text prompts [22, 50, 72], limiting their exploitation of structural guidance from reference images [14, 55, 66, 67, 70]. Parallel to this, **compositional generation and decomposition** require synthesizing coherent scenes from visual context. Although models like Echo-4o [59] and OmniGen [53, 57] address fixed configurations, they lack flexibility for decomposition [52, 56]. Most unified models are restricted to single-round inputs, often struggling with the long-horizon reasoning required for multi-element scenarios.

In contrast, Loom supports long-context multimodal tasks. Our framework scales to accommodate numerous visual inputs and enables diverse applications—ranging from high-fidelity style transfer and accuracy-critical virtual try-on [8, 9, 15, 63] to commercial model presentation, intricate multi-object arrangement, and reverse component extraction [6, 52, 60, 61, 66, 69], ensuring semantic consistency across extended sequences.

3. Methods

The overall architecture of Loom is shown in Fig 2. Our method consists of three components: (1) the Loom model design building on the Bagel backbone with task-specific training configurations; (2) a unified training paradigm with stepwise planning, sparse historical frame sampling, and temporal embeddings for coherent long-horizon multimodal generation; and (3) a curated 50K-sample interleaved dataset covering compositional generation, style transfer, and procedural tutorials.

3.1. Overall Architecture

Loom builds upon the Bagel backbone, which operates autoregressively over interleaved token sequences, where visual content is represented via a pre-trained VAE encoder and decoded through the rectified flow method. Following Bagel, we adopt full-parameter fine-tuning on the generative transformer backbone while keeping the ViT encoder frozen to preserve pre-trained semantic alignment. This allows Loom to unify text prediction and image generation within a single framework.

Multi-Modal Attention. Loom inherits Bagel's multi-modal attention (MMA) and noise isolation mechanism. To ensure stable autoregressive generation, we employ a noise-isolated causal masking strategy. The MMA formulation is defined as:

$$\mathrm{MMA}([c_T; c_Z; c_I]) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{1}$$

where $[c_T; c_Z; c_I]$ denotes the concatenation of three types of tokens: (1) c_T: text tokens, encoding semantic or instructional information. (2) c_Z: noised latent tokens, representing partially corrupted visual representations to be denoised. (3) c_I: image condition tokens, providing visual context or reference frames. Crucially, the masking is unidirectional: current noised tokens c_Z attend to history $[c_T; c_I]$ to retrieve context, while historical tokens are isolated from c_Z to prevent state contamination.

Interleaved Training Objects. We adopt the Bagel training objectives for interleaved synthesis, combining language modeling for text [1] tokens and rectified flow matching [27, 29] for image tokens. The overall training loss is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CE}} \cdot \mathcal{L}_{\text{CE}}^{\text{text}} + \mathcal{L}_{\text{MSE}}^{\text{image}} \tag{2}$$

Here, $\mathcal{L}_{\text{CE}}^{\text{text}}$ denotes the cross-entropy loss for next-token text prediction, and $\mathcal{L}_{\text{MSE}}^{\text{image}}$ denotes the mean squared error loss used for image denoising in latent space. The coefficient λ_{CE} controls the relative weight between textual and visual objectives. This unified objective allows Loom to jointly optimize heterogeneous interleaved tasks within a single autoregressive framework.

3.2. Interleaved Text Image Generation Tasks

Our training paradigm unifies heterogeneous interleaved tasks under a single autoregressive framework through three key algorithmic designs.

Planning-First Strategy. Standard interleaved models typically alternate between text and image generation steps ($T_1 \rightarrow I_1 \rightarrow T_2 \ldots$), where errors in early visual frames (I_1) propagate to subsequent text predictions (T_2), causing semantic drift [25] in long-horizon tasks. To mitigate this,

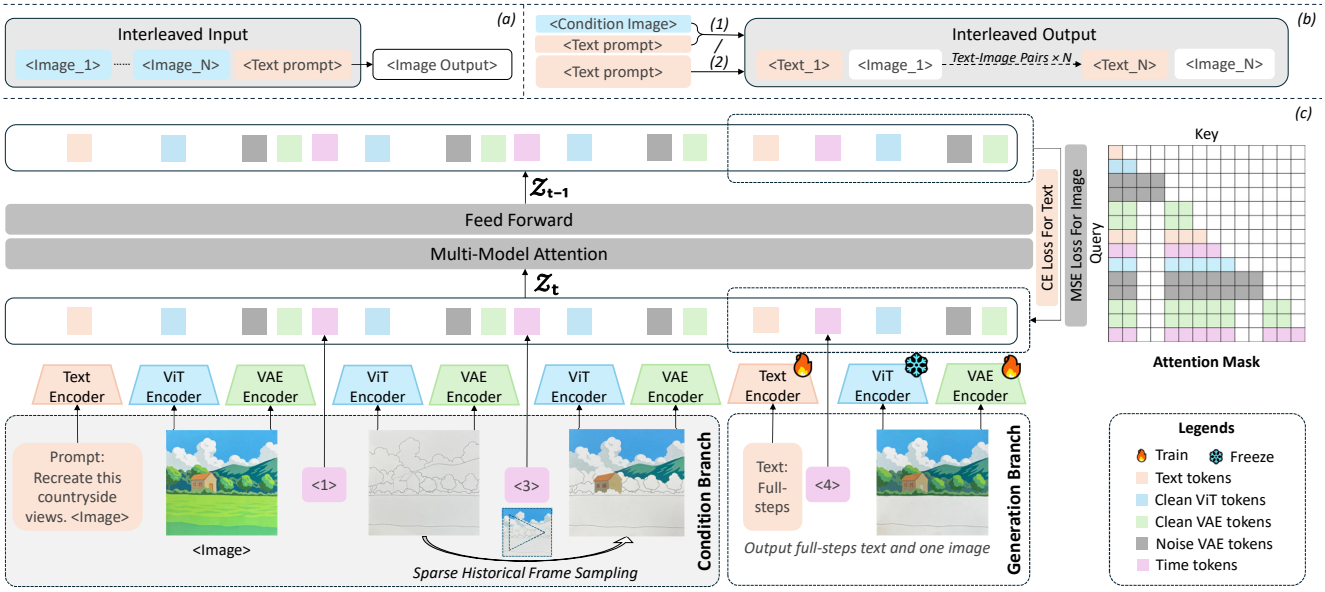

Figure 2. **(a)** An interleaved input paradigm with various conditional images and text prompts, producing a single image output. **(b)** Interleaved output paradigm, where the model takes either pure text instructions or mixed text-image guidance and generates multi-round, sequential text-image pairs. **(c)** Training and inference architecture for the interleaved output paradigm, focusing on case **(1)** where the input is a text-image guidance sequence and the output is continuous text-image pairs, exemplified by a step-by-step drawing tutorial. The pipeline contains a *condition branch*, which encodes sparse historical frames via ViT and VAE encoders to provide visual context, and a *generation branch*, which produces both full-step textual descriptions and the next image under the attention mask, ensuring alignment between global textual planning and incremental image rendering.

we propose a Planning-First Strategy that decouples high-level semantic reasoning from pixel-level synthesis.

Inspired by Chain-of-Thought reasoning [23, 40], Loom adopts a "plan-then-render" workflow. Given a instruction, the model first generates the entire formatted textual plan $\mathcal{P} = \{S_1, S_2, \ldots, S_N\}$ in a single autoregressive pass, decomposing the task into coherent substeps (e.g., `Step 1: prepare ingredients... Step N: final plating`) without visual interference. Once the plan is fixed, Loom enters a deterministic rendering loop: for each step t, the image I_t is generated conditioned on (1) the clean textual description S_t, (2) the global plan context $\mathcal{P}$, and (3) selected historical frames. For each frame I_t in an N-step sequence, we construct a training instance that predicts both the complete plan $\mathcal{P}$ and the current image I_t sequentially: the model first generates $\mathcal{P}$ autoregressively, then renders I_t conditioned on the corresponding step description extracted from $\mathcal{P}$. This multi-round alignment—where the same plan is trained with different visual states—ensures global coherence across all frames. As shown in Fig. 2(c), the attention mask ensures that noised VAE tokens (gray blocks) are isolated from previous steps; only clean historical frames (green blocks) condition the generation. This noise isolation design prevents the model from learning spurious correlations with intermediate diffusion states, preventing noisy context from contaminating the logical progression of the text.

Temporal Embeddings. To further stabilize long-horizon interleaved generation, we incorporate a learnable time embedding that explicitly encodes the relative position of each visual frame within the planned generation sequence. For a frame at step t, we add a dedicated vector $\mathbf{e}_t$ to the input sequence via broadcasting:

$$\tilde{\mathbf{x}}_t = [c_T; c_{I,Z}^{(t-1)}; \mathbf{v}_t] + \mathbf{e}_t, \tag{3}$$

where $\mathbf{v}_t$ denotes the visual tokens of the t-th frame, c_T represents the textual tokens, and $c_{I,Z}^{(t-1)}$ are image-condition tokens from step $t-1$. $\mathbf{e}_t \in \mathbb{R}^d$ is a learnable temporal embedding for step t, which is broadcasted and added to all tokens in $\tilde{\mathbf{x}}_t$ to globally inform the model of the current frame's position in the generation sequence.

Sparse Historical Frame Sampling. In interleaved text–image generation, natively conditioning each step on a fixed-length sliding window of recent frames tends to overfit local patterns, making the model insensitive to the global textual plan. To retain both efficiency and global awareness, we adopt a sparse historical frame sampling strategy. Instead of relying on complex heuristics, our approach selects a small but temporally diverse subset of frames by sampling from the entire history at evenly distributed intervals.

This method ensures the model receives a balanced mix of contextual information—from early frames that establish the overarching structure to recent frames that detail the latest changes. For each selected frame, visual features are extracted, augmented with a learnable temporal embedding representing its absolute step index, and injected into the multimodal attention block. This allows the model to integrate long-range scene structure and short-term visual detail with negligible overhead

3.3. Entity Tokens for Compositional Control

Standard text prompts lack explicit handles for individual objects, making fine-grained control difficult in compositional and multi-image reasoning tasks. We introduce learnable entity tokens (e.g., ⟨model⟩, ⟨garment⟩) that serve as semantic anchors, immediately followed by detailed descriptions. For example: "a ⟨model⟩ wearing a ⟨garment⟩ [red floral dress]". This structure allows the model to bind entity tokens to specific visual concepts during training, enabling both forward generation and reverse decomposition. Crucially, the clean object isolation in our dataset provides implicit spatial supervision, teaching the model to restrict the influence of entity tokens strictly to their corresponding pixel regions to prevent attribute leakage. This effectively renders the gradient updates spatially sparse without requiring manual masks.

To further enhance controllability, we leverage the multimodal classifier-free guidance (CFG) [17] capability of the Bagel backbone, extending beyond traditional text-only guidance. When combined with standard tokens (e.g., ⟨upper⟩, ⟨body⟩) that are consistently used across training and inference, Loom can localize guidance signals to specific objects, selectively applying multimodal constraints to targeted regions. This synergy between structured entity grounding and multi-modal CFG yields precise, interpretable, and robust compositional control across both synthesis and decomposition tasks.

3.4. Interleaved Dataset Construction

Creating a robust interleaved generator requires data that excels in both temporal logic and text-image alignment. As shown in Fig 3, we construct a high-fidelity 50K-sample dataset through a rigorous multi-stage pipeline spanning three distinct task categories.

Compositional Generation and Decomposition. We curate a logically complex subset from Echo-4o[59] and our internal imagery (11k, generated by Gemini 2.5 Flash Image [11]), focusing on multi-object interactions. Unlike standard datasets, we generate paired bi-directional instructions: forward commands for integrating entities into unified layouts, and reverse commands for decomposing scenes into constituent parts.

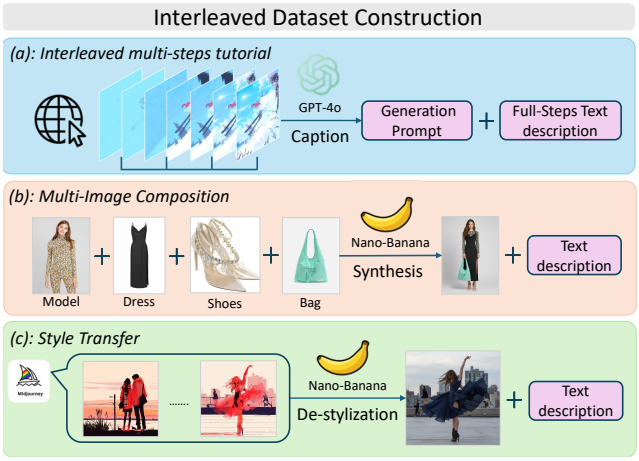

Figure 3. Interleaved dataset construction: (a) Blogs and videos are collected, and 4–6 frames are uniformly sampled, manually verified, and captioned by GPT-4o with generation prompts and stepwise captions. (b) Multi-image composition combines models, objects, and scenes via Nano-Banana [11] with textual descriptions. (c) Style transfer uses Promptsref [39] website images; Nano-Banana performs de-stylization to obtain realistic images and corresponding prompts.

Style Transfer Pairs. We construct a paired stylization dataset using 12k high-aesthetic references from Promptsref [39]. To enforce content-structure disentanglement, we employ a "De-stylization" pipeline using Gemini 2.5 Flash Image (Nano-Banana) to generate photorealistic content counterparts for each stylized reference. This results in perfectly aligned (Content, Style, Output) triplets, enabling the model to learn precise style injection while preserving structural fidelity under supervised conditions.

Sequential Procedural Tutorials. We focus on long-horizon logic in cooking and creative arts. For cooking, we extract a high-alignment subset (14k) from RecipeGen[62], filtering for sequences with $\geq$ 3 steps where visual state changes strictly follow textual instructions. For painting, we build a specialized 8k dataset from iPad-based screen recordings, manually filtered to reflect human cognitive progression (sketch $\rightarrow$ color $\rightarrow$ detail). We also use GPT-4o to refine stepwise captions for causal consistency from MakeAnything(5k sketch painting). This tiered filtration ensures that every frame transition in our training data serves as a valid logical progression for the planner.

4. Experiment

4.1. Set up

Implementation details. We employ the supervised fine-tuning (SFT) method; we set the learning rate 2.5×10^{-5} with a constant scheduler, zero weight decay, and gradient norm clipping at 1.0. Optimization uses AdamW with a CE to MSE loss weight ratio of $0.25 : 1$. The total training

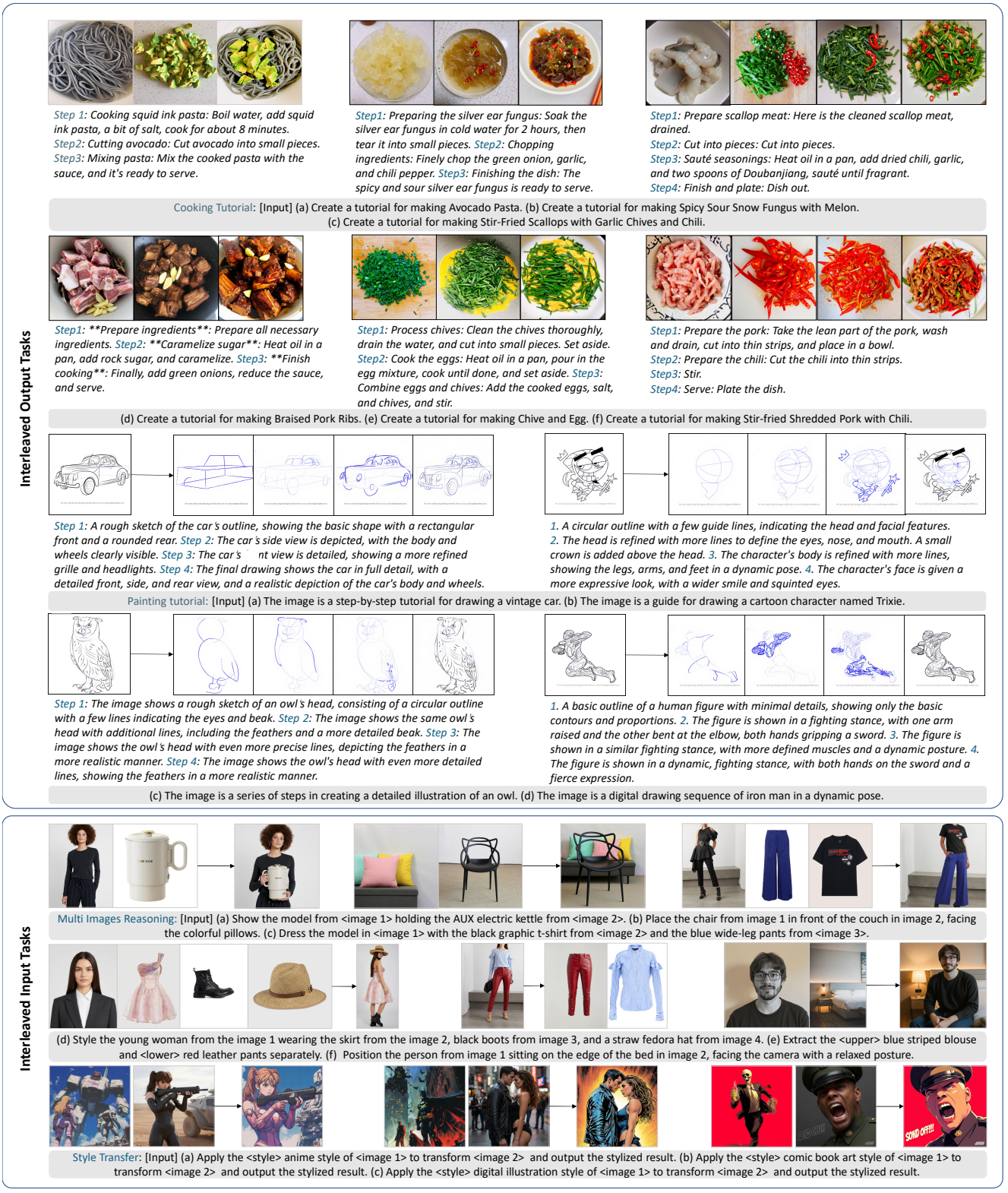

Figure 4. More generation results of Loom in interleaved tasks. The top rows show interleaved output tasks, including text-to-interleaved cooking tutorials and image-to-interleaved painting tutorials. The bottom rows depict interleaved input tasks, covering (1) procedural generation, (2) compositional generation and decomposition, and (3) style transfer from a given reference image.

steps are set as follows: 16,000 steps for the compositional generation (long-text, multi-image reasoning) task, 10,000 steps for style transfer, 20,000 steps for recipe-oriented text–image interactive generation, and 5,000 steps for interleaved painting tutorials. The batch size is fixed at 4. We employ an Exponential Moving Average (EMA) ratio 0.999 and sequence length per rank in the range of $(25K, 29K)$ tokens, with a maximum context window of 29K. All generations are performed at a resolution between $(512, 1024)$ pixels on the short and long sides, with unconditional image resolution between $(378, 980)$ pixels.

Benchmarks. For the multi-image reasoning task, we adopt the OmniContext benchmark, which evaluates the ability to integrate multiple visual references into coherent outputs by measuring scene completeness, object fidelity, and semantic consistency.

Metrics. We evaluate interleaved generation in two settings: image-conditioned and text-only. Performance is measured across three dimensions: *visual continuity* (temporal coherence between frames), *textual continuity* (logical flow of descriptions), and *cross-modal alignment* (consistency between images and instructions). For text-to-interleaved generation, we also report the CLIP [41] (ViT-L/14) score to quantify prompt–image alignment. Human ratings are compared with GPT-based scores to validate metric reliability.

Baseline methods. For the text–image interleaved generation task, we compare our approach with Anole. We also adapt unified models such as Janus-Pro and Bagel by constructing a multi-turn dialogue framework to support interleaved outputs and by providing extended text prompts to enable long context, continuous generation beyond their native single-turn editing capabilities. In addition, we include a proprietary large multimodal model, Doubao, as a reference baseline due to its support for high-quality interactive text–image generation.

4.2. Quantitative Comparison

Table 1. Text-to-interleaved results using GPT-4o(G) and Human(H) scoring. Coh. = Temporal Coherence; Ins. = Instruction Following; Con. = Narrative Consistency; CLIP = CLIP Score; G = GPT-4o score; H = Human score.

Model	Coh.(G | H)↑	Ins.(G | H)↑	Con.(G | H)↑	CLIP↑
Doubao [13]	**4.35 | 4.10**	**4.25 | 4.05**	**4.95 | 4.65**	0.250
Bagel [12]	1.40 | 1.25	1.55 | 1.05	–	0.217
Janus-Pro [5]	1.05 | 1.10	1.10 | 1.00	–	0.105
Anole [7]	1.55 | 1.05	1.35 | 1.05	1.95 | 1.35	0.219
Loom (Ours)	**4.25 | 4.15**	**3.75 | 3.35**	**4.70 | 4.30**	**0.269**

Task 1: Interleaved generation. We evaluate two sub-settings: text-to-interleaved (Table 1) and image-to-interleaved (Table 2). Compared to the baseline Bagel,

Table 2. Image-to-interleaved results using GPT-4o and human scoring. Coh. = Temporal Coherence; Ref. = Reference Faithfulness; Ali. = Semantic Alignment; G = GPT-4o score; H = Human score.

Model	Coh. (G | H)↑	Ref. (G | H)↑	Ali. (G | H)↑
Doubao [13]	2.05 | 2.15	2.65 | 2.65	**3.55 | 3.85**
Bagel [12]	1.25 | 1.00	1.15 | 1.55	–
Loom (Ours)	**3.15 | 3.65**	**3.85 | 4.15**	**3.15 | 2.95**

Loom achieves an average relative improvement of over 50% in text-to-interleaved generation and around 44% in image-to-interleaved generation, while outperforming the strongest open-source interleaved model Anole by over 51% in the text setting. *(Con. and Ali. are not reported for baseline model Bagel and Janus-Pro that do not produce paired text-image guidance in the given setting.)*

In the text-to-interleaved setting, extending Bagel with our multi-turn framework enables continuous stepwise generation from scratch, where the native model cannot. Across metrics, Loom far surpasses Anole and all single-turn unified baselines, and in image-to-interleaved task even exceeds the proprietary model Doubao in some metrics, evidencing strong overall capability. In the image-to-interleaved setting, Loom achieves the highest average scores, balancing instruction adherence and faithful visual preservation.

Qualitative comparisons in Figure 5 show that open-source single-turn unified models quickly lose logical progression in later frames, while Hunyuan Image Edit [2] fail to preserve details. Anole struggles with text–image alignment; Doubao produces visually appealing and well-aligned results but still exhibits logical breaks, whereas Loom maintains superior alignment, coherence, and near-perfect adherence to reference images in painting tasks.

Table 3. Multi-image reasoning performance on the OmniContext benchmark. Scores are categorized by scene type (MULTIPLE and SCENE) and the composition of "Character" (Char.), "Object" (Obj.), and both (Char.+Obj.). Higher average scores (Avg) indicate better performance. The benchmark's SINGLE category is excluded, as it is not relevant to our multi-image task.

Model	MULTIPLE			SCENE			Avg↑
	Char.	Object	Char.+Obj.	Char.	Object	Char.+Obj.	
GPT-4o [38]	**9.07**	**8.95**	**8.54**	**8.90**	**8.44**	**8.60**	**8.75**
UNO [56]	2.54	6.51	4.39	2.06	4.33	4.37	4.03
Bagel [12]	5.17	6.64	6.24	4.07	5.71	5.47	5.55
OmniGen [57]	5.65	5.44	4.68	3.59	4.32	5.12	4.8
OmniGen2 [53]	7.11	7.13	7.45	6.38	6.71	7.04	6.97
Echo-4o [59]	8.07	7.50	8.29	8.62	8.00	8.08	8.09
Loom(Ours)	**8.09**	**7.62**	**8.25**	**8.67**	**7.95**	**8.23**	**8.13**

Task 2: Compositional Generation (multi-image reasoning). For this task, performance is directly reported using the OmniContext benchmark in Table 3.

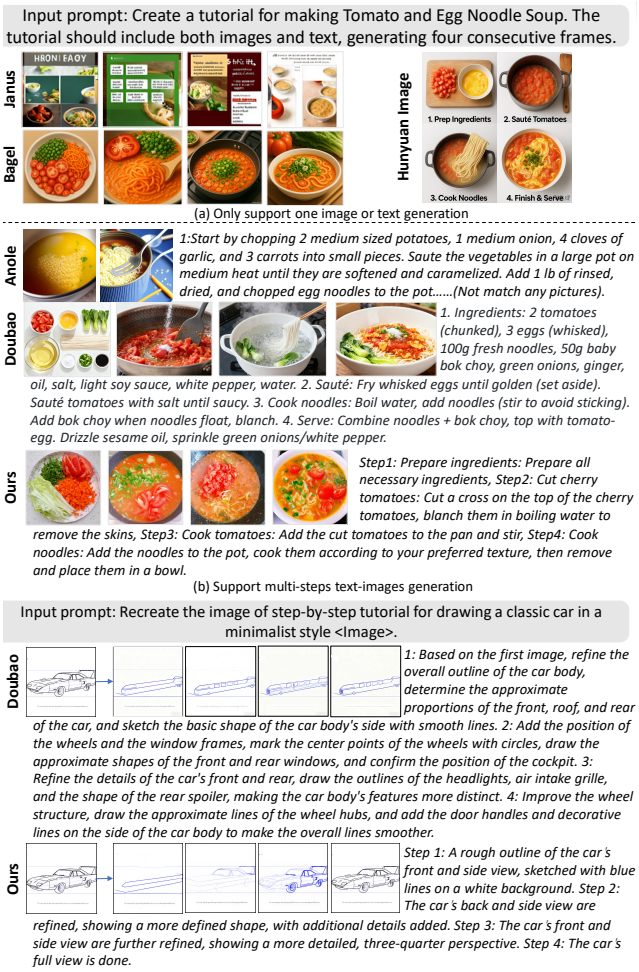

Figure 5. Comparison results. (a) Unified models such as Bagel and Janus-Pro only support single-round input–output generation. (b) Interleaved models, including Anole and Doubao support multi-step text–image generation; we also compared closed-source Doubao with our method.

4.3. Qualitative Evaluation

More qualitative generation results are shown in Figure 4.

4.4. Ablation

We perform ablation studies exclusively on the image-to-interleaved generation setting (painting tutorials). More ablation results are in Figure 6. The ablation results in Table 4 reveal contributions from all components. Removing any single module leads to a clear drop across at least two metrics, confirming their complementary roles: textual guidance provides fine-grained control over semantic alignment, reference sampling strengthens temporal coherence via richer visual context, and time embedding is critical for maintaining consistent progression across frames. Compared to the baseline model Bagel, the full system improves temporal coherence by +38%, reference faithfulness by +54%, underscoring the necessity of all components.

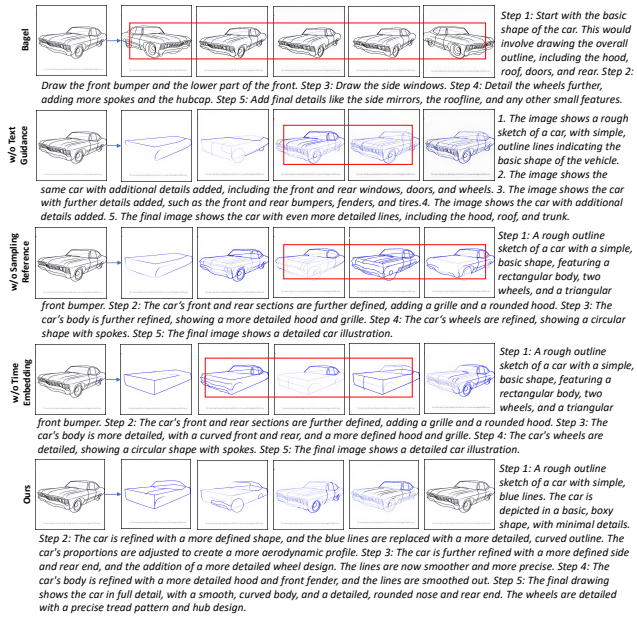

Figure 6. Ablation study results. From top to bottom: (1) baseline (Bagel), (2–4) removing each of the three modules: stepwise textual guidance, reference sampling, and time embedding and (5) our full model Loom. Red rectangles highlight errors.

Table 4. Ablation Study for Painting Tutorials (Image-to-Interleave) with GPT-4o Evaluation. Coherence. = Temporal Coherence; Reference. = Reference Faithfulness; Alignment. = Semantic Alignment.

Method	Coherence.↑	Reference.↑	Alignment.↑
Baseline	1.25	1.15	–
w/o Time Embedding	2.55	2.35	2.85
w/o Stepwise Prompt	2.15	2.95	2.35
w/o Reference Sampling	1.45	1.25	2.05
Full	**3.15**	**3.85**	**3.15**

5. Conclusion

We present **Loom**, a unified diffusion–transformer framework for interleaved text–image generation across style transfer, compositional synthesis, and procedural tutorials. Built by extending the Bagel pre-trained unified model from single-turn inputs to handle complex N-to-M interleaved sequences, Loom supports interleaved procedural sequencing, structured compositional manipulation, and reference-guided appearance transfer. Experiments on a 50K interleaved dataset demonstrate that Loom delivers superior compositionality, temporal coherence, and text–image alignment, significantly outperforming open-source baselines and setting a new benchmark for coherent interleaved multimodal generation.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 3

[2] 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. 7

[3] Bowei Chen, Yifan Wang, Brian Curless, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. Inverse painting: Reconstructing the painting process, 2024. 3

[4] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. 1

[5] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. 7

[6] Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, Runsheng Yu, Yiming Yu, Liehao Zou, Hang Li, Lu Lu, Yuxuan Wang, and Yonghui Wu. Seed-x: Building strong multilingual translation llm with 7b parameters, 2025. 3

[7] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation, 2024. 3, 7

[8] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild, 2024. 3

[9] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models, 2025. 3

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, and Inderjit Dhillon et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 2

[11] Google Deepmind. Nano-banana. https://deepmind.google/models/gemini-image/pro/, 2025. Accessed: 2025-11-14. 5

[12] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1, 2, 7

[13] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xuanda Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Xin Xia, Xuefeng Xiao, Zhonghua Zhai, Xinyu Zhang, Qi Zhang, Yuwei Zhang, Shijia Zhao, Jianchao Yang, and Weilin Huang. Seedream 3.0 technical report, 2025. 2, 7

[14] Yan Gong, Yiren Song, Yicheng Li, Chenglin Li, and Yin Zhang. Relationadapter: Learning and transferring visual relation with diffusion transformers. *arXiv preprint arXiv:2506.02528*, 2025. 3

[15] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Chuang Zhang, and Jiaming Liu. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. *arXiv preprint arXiv:2501.15891*, 2025. 3

[16] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention, 2024. 3

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 5

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 3

[19] Chengyou Jia, Xin Shen, Zhuohang Dang, Zhuohang Dang, Changliang Xia, Weijia Wu, Xinyu Zhang, Hangwei Qian, Ivor W. Tsang, and Minnan Luo. Why settle for one? text-to-imageset generation and evaluation, 2025. 3

[20] 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, 2025. 3

[21] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3

[22] Mingkun Lei, Xue Song, Beier Zhu, Hao Wang, and Chi Zhang. Stylestudio: Text-driven style transfer with selective control of style elements, 2025. 3

[23] Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision language reasoning, 2025. 4

[24] Qi Li, Runpeng Yu, Haiquan Lu, and Xinchao Wang. Every step counts: Decoding trajectories as authorship fingerprints of dllms. *arXiv preprint arXiv:2510.05148*, 2025. 3

[25] 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, 2025. 3

[26] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, Yatian Pang, and Li Yuan. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation, 2025. 1

[27] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. 3

[28] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport, 2022. 2

[29] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. 3

[30] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[31] Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.

[32] Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025.

[33] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6018–6026, 2025. 3

[34] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2025. 1

[35] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 3

[36] Yue Ma, Zexuan Yan, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, et al. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *arXiv preprint arXiv:2509.16630*, 2025. 3

[37] John Nguyen, Marton Havasi, Tariq Berrada, Luke Zettlemoyer, and Ricky T. Q. Chen. Oneflow: Concurrent mixed-modal and interleaved generation with edit flows, 2025. 3

[38] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, and Adam Perelman et al. Gpt-4o system card, 2024. 2, 7

[39] Promptsref. Promptsref. https://promptsref.com/zh, 2025. Accessed: 2025-11-14. 5

[40] Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision, 2025. 4

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 7

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 3

[43] Yiren Song, Danze Chen, and Mike Zheng Shou. Layertracer: Cognitive-aligned layered svg synthesis via diffusion transformer. *arXiv preprint arXiv:2502.01105*, 2025. 3

[44] Yiren Song, Cheng Liu, and Mike Zheng Shou. Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation, 2025. 3

[45] Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data, 2025. 3

[46] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. 2

[47] Paints-Undo Team. Paints-undo github page, 2024. 3

[48] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, Hongsheng Li, Yu Qiao, and Jifeng Dai. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer, 2024. 3

[49] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 2

[50] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation, 2024. 3

[51] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. 2

[52] Zitong Wang, Hang Zhao, Qianyu Zhou, Xuequan Lu, Xiangtai Li, and Yiren Song. Diffdecompose: Layer-wise decomposition of alpha-composited images via diffusion transformers. *arXiv preprint arXiv:2505.21541*, 2025. 3

[53] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation, 2025. 3, 7

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

[55] Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning, 2025. 3

[56] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 3, 7

[57] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 3, 7

[58] 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, 2025. 1, 2

[59] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, Conghui He, and Weijia Li. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation, 2025. 3, 5, 7

[60] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2024. 3

[61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3

[62] Ruoxuan Zhang, Hongxia Xie, Yi Yao, Jian-Yu Jiang-Lin, Bin Wen, Ling Lo, Hong-Han Shuai, Yung-Hui Li, and Wen-Huang Cheng. Recipegen: A benchmark for real-world recipe image generation, 2025. 3, 5

[63] Xuanpu Zhang, Dan Song, Pengxin Zhan, Tianyu Chang, Jianhao Zeng, Qingguo Chen, Weihua Luo, and Anan Liu. Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training, 2024. 3

[64] Xiaoyan Zhang, Zechen Bai, Haofan Wang, and Yiren Song. Sigma: Selective-interleaved generation with multi-attribute tokens. *arXiv preprint arXiv:2602.07564*, 2026. 2

[65] Yu Zhang, Jialei Zhou, Xinchen Li, Qi Zhang, Zhongwei Wan, Duoqian Miao, Changwei Wang, and Longbing Cao. Enhancing text-to-image diffusion transformer via split-text conditioning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2

[66] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078, 2024. 3

[67] Yuxuan Zhang, Lifu Wei, Qing Zhang, Yiren Song, Jiaming Liu, Huaxia Li, Xu Tang, Yao Hu, and Haibo Zhao. Stable-makeup: When real-world makeup transfer meets diffusion model. *arXiv preprint arXiv:2403.07764*, 2024. 3

[68] Yu Zhang, Jingyi Liu, Yiwei Shi, Qi Zhang, Duoqian Miao, Changwei Wang, and Longbing Cao. Markovian scale prediction: A new era of visual autoregressive generation. *arXiv preprint arXiv:2511.23334*, 2025. 2

[69] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025. 3

[70] Yuxuan Zhang, Qing Zhang, Yiren Song, Jichao Zhang, Hao Tang, and Jiaming Liu. Stable-hair: Real-world hair transfer via diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10348–10356, 2025. 3

[71] Yu Zhang, Jingyi Liu, Feng Liu, Duoqian Miao, Qi Zhang, Kexue Fu, Changwei Wang, and Longbing Cao. Adaptive visual autoregressive acceleration via dual-linkage entropy analysis. *arXiv preprint arXiv:2602.01345*, 2026. 2

[72] Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm, 2024. 3