SCALING UNIFIED MULTIMODAL MODELS TO LONG-HORIZON INTERLEAVED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models (UMMs) provide an interface for interleaved text-image generation, where generated visual states can directly inform subsequent reasoning and synthesis. However, existing systems are still primarily demonstrated on short trajectories or simplified visual content. We develop a large UMM natively trained on interleaved multimodal data, with strong capabilities in image synthesis, text rendering, complex composition, and multimodal planning, and study how to extend these native capabilities to substantially longer autoregressive trajectories. The main challenge is that generated images continuously accumulate in context: irrelevant or imperfect visual states interfere with future synthesis, while previously established characters and scenes must still be recovered after long intervals. We introduce a visual-memory framework that maintains persistent visual references, selectively retrieves scene-relevant historical states while preserving the complete semantic history, and uses few-step self-rollout to adapt the model to imperfect generated histories. We further introduce LongMIG-Bench for evaluating long-horizon text-to-multi-image and image-to-multi-image generation. Our model achieves high-fidelity long-horizon interleaved generation with strong character and scene consistency, demonstrating how visual memory can scale native interleaved generation to substantially longer trajectories.

1 INTRODUCTION

Unified multimodal models (UMMs) are evolving from systems that separately understand or generate visual content into general-purpose models that can reason and generate across text and images within a single autoregressive trajectory (Team, 2024; Deng et al., 2025; Wu et al., 2025; Wang et al., 2024a; 2026; Xie et al., 2025b). Native interleaving provides a natural formulation for long-form multimodal generation: generated visual states can directly inform subsequent reasoning and synthesis, allowing the model to continuously update its plan based on previously generated text and images rather than reducing each step to an independent prompt. It also provides a natural interface for learning from sequential visual data such as videos, comics, storyboards, tutorials, and other multimodal sequences. These properties make native interleaved generation a promising paradigm for applications including illustrated stories, visual design, tutorials, and multi-shot content creation.

Despite this potential, existing interleaved systems are still primarily demonstrated on relatively short trajectories or simplified visual content (Yang et al., 2025; Chen et al., 2026a). In this work, we train a large UMM natively on mixed text-image trajectories, where visual outputs remain within the shared multimodal sequence and can directly condition subsequent reasoning and synthesis. The resulting model supports high-fidelity image synthesis, text rendering, complex composition, and multimodal planning within a single autoregressive trajectory. Our goal is not merely to generate more images in sequence, but to preserve these native capabilities over substantially longer horizons.

However, native interleaved training alone does not resolve the challenges that emerge as the trajectory becomes long. Generated images themselves become part of the future context, and every image introduces a dense visual state that may influence later synthesis. Some states remain essential, such as the appearance of a recurring character or an earlier view of a scene, while others become redundant, irrelevant, or imperfect. Conditioning every image on the complete visual history therefore leads to continuously growing visual context and increasing interference, while relying only on recent images may discard information needed for long-range revisits (Ye et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **High-fidelity long-horizon interleaved generation across diverse visual domains.** Our model supports photorealistic, stylized, text-rich, and design-oriented generation while maintaining visual consistency over extended trajectories.

Recent work has similarly shown that retaining more visual context does not necessarily improve long-horizon generation (Chen et al., 2026a). Meanwhile, teacher-forced training conditions on clean ground-truth histories, whereas autoregressive inference must continue from the model's own outputs, allowing small visual errors to accumulate over time (Huang et al., 2025; Chen et al., 2024).

We address these challenges through an explicit visual-memory framework. The model first forms persistent visual references for information that should remain stable throughout the trajectory. Before each image generation, its understanding pathway predicts the current scene state and selectively retrieves relevant historical visual states. Importantly, this selection applies only to dense generative visual representations, while the complete semantic history remains available for global reasoning. We further introduce few-step self-rollout adaptation, which exposes the model during post-training to its own generated histories while retaining ground-truth targets for supervision. These mechanisms reduce unnecessary historical visual conditioning while preserving information required for long-range consistency.

To evaluate this setting, we introduce LongMIG-Bench for long-horizon text-to-multi-image and image-to-multi-image generation. The benchmark contains recurring characters, scene transitions, and long-range revisits, requiring models to preserve and recover visual information across multiple intervening generations. Experiments show that our model achieves high-fidelity long-horizon interleaved generation with strong character and scene consistency, while controlled analyses demonstrate the complementary roles of persistent visual references, scene-aware retrieval, and self-rollout adaptation. Our memory design also keeps the dense historical visual context used for image synthesis compact, enabling longer and more efficient autoregressive trajectories.

Our contributions are threefold:

- We develop a large UMM natively trained for interleaved text-image generation and study how to scale its image synthesis, text rendering, complex composition, and multimodal planning capabilities to long autoregressive trajectories.
- We develop a visual-memory framework that combines persistent visual references, scene-aware historical retrieval, and few-step self-rollout adaptation, while separating complete semantic history from the dense visual states directly used for image synthesis.
- We introduce LongMIG-Bench to evaluate long-horizon interleaved generation and demonstrate high-fidelity generation with strong long-range character and scene consistency.

2 Related Work

2.1 Unified Multimodal and Long-Horizon Generation

Recent unified multimodal models (Diao et al., 2026; Xie et al., 2025b) integrate visual understanding and generation within a single architecture, enabling text and images to be processed and generated in flexible interleaved sequences. Unlike agent-based pipelines (Wang et al., 2024b; Zhang et al., 2026) that reduce each generation step to an independent image prompt, native interleaved models retain generated visual states within the same autoregressive trajectory, enabling generation and planning to continuously condition on previous multimodal outputs. Representative models such as Janus (Wu et al., 2025), BAGEL (Deng et al., 2025), Show-o (Xie et al., 2025a), SEED-Story (Yang et al., 2025), Wan-Weaver (Xing et al., 2026), and ChatUMM (Dai et al., 2026) demonstrate increasingly strong capabilities in multimodal understanding, image generation, and interleaved text-image generation. These advances establish that unified models can produce multimodal outputs within a single generative trajectory.

Our focus is on how this capability scales to substantially longer horizons. As generated images accumulate in the context, models must preserve visual information across distant steps while controlling an increasingly large history. Related problems have been studied in long-context modeling through selective state retention and KV-cache compression (Xiao et al., 2024; Li et al., 2024). More closely, UniLongGen (Chen et al., 2026a) shows that retaining the complete visual history can degrade long-form generation and introduces training-free context curation to remove less relevant historical states. Our work shares the motivation of avoiding indiscriminate full-history conditioning, but instead organizes visual history as explicit memory: the model preserves the full semantic trajectory for global reasoning while selectively exposing task-relevant visual states to image generation. We therefore study visual-memory management as a mechanism for scaling native interleaved generation to long multimodal trajectories.

2.2 Visual Memory and Self-Conditioned Generation

Recent multimodal models have begun to treat generated images as intermediate reasoning states. Works on generated visual thoughts and multimodal chain-of-thought (Chern et al., 2025; Qin et al., 2026; Chen et al., 2026b; Yin et al., 2026; Hu et al., 2026) show that visual outputs can support reasoning, refinement, and test-time computation. Our setting differs in that generated visual states are used as *persistent memory* rather than transient intermediate reasoning. The model autonomously creates reusable character references because their information will be needed repeatedly later in the trajectory. This is also related to subject-driven generation (Ruiz et al., 2023; Chen et al., 2023; Ye et al., 2023), but our references are generated by the model itself rather than supplied externally.

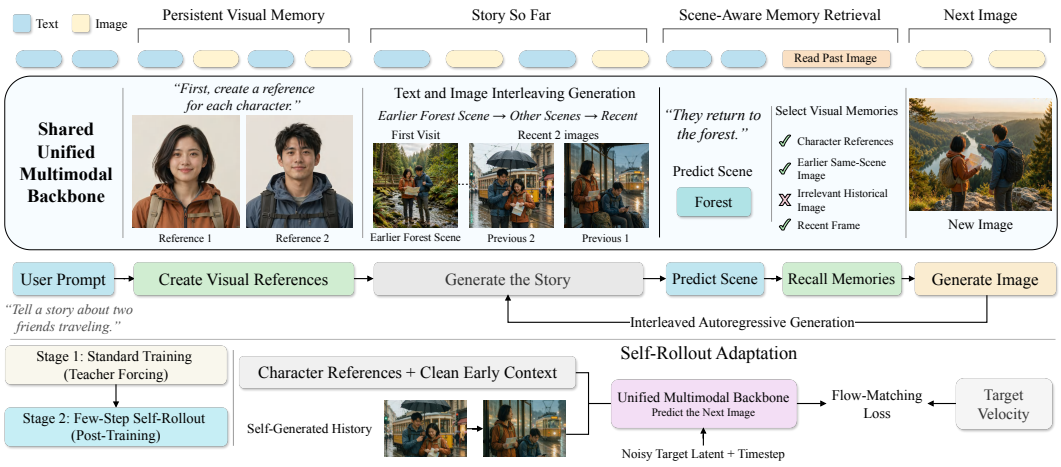


Figure 2: **Overview of our long-horizon interleaved generation framework.** Built on our unified multimodal backbone, the model constructs persistent visual references, retrieves scene-relevant history during autoregressive generation, and uses few-step self-rollout for memory adaptation.

Long-horizon generation amplifies the train–test mismatch of autoregressive generation. Teacher-forced training conditions on ground-truth histories, whereas inference must continue from the model's own imperfect outputs. Rollout-based approaches such as Self-Forcing (Huang et al., 2025), Self-Forcing++ (Cui et al., 2026), and Rolling Forcing (Liu et al., 2026) address exposure gaps in autoregressive video generation. We study the analogous problem for interleaved multimodal sequences, using few-step self-rollout to expose the model to self-generated visual histories while retaining ground-truth targets for supervision. These directions motivate our view of long-horizon interleaved generation as a visual-memory problem: models must determine what visual information should persist, what should be retrieved at each step, and how to continue from imperfect history.

3 Method

3.1 Overview

We consider long-horizon interleaved generation with our unified multimodal model, natively trained on mixed text-image trajectories. During training, text and image outputs are represented within the same evolving multimodal sequence, allowing subsequent reasoning and synthesis to condition directly on preceding multimodal states. Given an instruction x, the model produces

$$Y = (y_1, y_2, \ldots, y_T), \tag{1}$$

where each y_t is either text or an image. Let τ_k denote the position of the k-th generated image I_k in the multimodal sequence. Native interleaved generation then conditions on the preceding multimodal history,

$$p_\theta(I_k \mid x, Y_{<\tau_k}). \tag{2}$$

Native interleaved training provides the model with the ability to reason and generate within a shared evolving multimodal trajectory, but does not by itself resolve the accumulation of dense visual states at long horizons. As generation proceeds, each previous image contributes a dense visual state, although only a subset may remain useful for the current synthesis.

We formulate long-horizon interleaved generation as a visual-memory problem. For each historical image I_j, the model maintains two representations: a semantic representation h_j^{sem} for multimodal understanding and a generative visual representation m_j for image synthesis. We preserve the semantic trajectory for global reasoning, while controlling which historical generative representations condition each new image. This allows the model to reason over the full multimodal history without requiring every previous image to participate in visual generation. Our framework has three components: persistent visual memory for stable information, scene-aware retrieval for selecting relevant historical states, and self-rollout adaptation for training on imperfect autoregressive histories.

3.2 PERSISTENT VISUAL MEMORY

Long-form generation repeatedly revisits information that should remain stable despite changes in pose, viewpoint, lighting, or composition. Character identity is a representative example. Relying only on arbitrary previously generated story images makes such information increasingly difficult to recover as the trajectory grows, since those images may themselves contain pose-specific details or generation errors.

We therefore allow the model to explicitly construct reusable visual references before long-form generation. Given instruction x, the model first generates a multimodal reasoning trajectory containing a set of visual anchors

$$\mathcal{A} = \{A_1, \ldots, A_M\}, \tag{3}$$

followed by the interleaved output trajectory,

$$\underbrace{\langle \text{think} \rangle\, r^{\text{text}}, A_1, \ldots, A_M\, \langle /\text{think} \rangle}_{\text{visual memory formation}} \quad \underbrace{\langle \text{answer} \rangle\, t_1, I_1, \ldots, t_N, I_N\, \langle /\text{answer} \rangle}_{\text{interleaved generation}}. \tag{4}$$

The anchors are generated by the model itself rather than provided externally and are trained using the model's native image-generation objective. Once generated, their clean generative representations are retained throughout the trajectory. They therefore provide stable visual references that can be reused at distant generation steps instead of repeatedly reconstructing identity from imperfect intermediate outputs.

3.3 SCENE-AWARE VISUAL MEMORY RETRIEVAL

Persistent anchors preserve globally important information, but ordinary historical images are not equally useful for every generation step. A fixed recent window preserves local continuity but fails when the model revisits an earlier scene, whereas full-history conditioning introduces an increasing number of irrelevant visual states. We therefore use the model's understanding pathway to determine which historical images should directly participate in image synthesis.

Before generating I_k, the model predicts a discrete scene state

$$s_k \in \{0, 1, \ldots, S-1\}, \tag{5}$$

from the instruction and preceding multimodal trajectory. During training, scene prediction is supervised using scene annotations; during inference, s_k is produced autoregressively by the model itself. The predicted state acts as an interface between multimodal understanding and visual generation: the model first determines the semantic state of the upcoming image and then retrieves visual memories associated with that state.

For generation step k, we construct the accessible generative visual memory as

$$\mathcal{M}_k = \mathcal{A} \cup \mathcal{R}_k \cup \mathcal{F}_k \cup \mathcal{L}_k, \tag{6}$$

where $\mathcal{A}$ contains persistent visual anchors, $\mathcal{R}_k$ contains the most recent K_r visual states regardless of scene, $\mathcal{F}_k$ contains early memories from the current scene, and $\mathcal{L}_k$ contains its most recent memories. For an ordinary historical image I_j, visibility to the generation of I_k can be written as

$$g(k,j) = \mathbb{I}[0 < k - j \le K_r] \;\vee\; \Big(\mathbb{I}[s_j = s_k] \wedge \big[\mathbb{I}[\rho(j) < N] \vee \mathbb{I}[\rho(k) - M \le \rho(j) < \rho(k)]\big]\Big), \tag{7}$$

where $\rho(j)$ denotes the occurrence index of I_j within its scene. The recent window maintains short-range continuity, while scene-aware retrieval recovers both establishing and recent views when a scene is revisited after a long interval.

Crucially, this routing is applied only to historical generative visual representations. Images excluded from $\mathcal{M}_k$ remain accessible through their semantic representations h_j^{sem}, allowing the understanding pathway to reason over the complete trajectory. We implement the retrieval rule by controlling attention from current image-generation queries to historical visual-memory keys, so excluded visual blocks need not participate in dense generative attention. The resulting generative visual memory therefore remains small even as the semantic trajectory continues to grow.

3.4 Self-Rollout Adaptation

Memory formation and retrieval determine what information is available to the generator, but long autoregressive trajectories introduce an additional train–test mismatch. Standard teacher-forced training conditions each image on clean ground-truth history,

$$I_1^*, I_2^*, \ldots, I_{k-1}^* \rightarrow I_k^*, \tag{8}$$

whereas inference conditions on the model's own previous generations,

$$\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_{k-1} \rightarrow \hat{I}_k. \tag{9}$$

Small deviations in identity, color, structure, or composition therefore become part of the subsequent visual context and may accumulate over many generation steps.

We reduce this exposure gap with a lightweight self-rollout stage. Starting from the teacher-forced model, we generate short autoregressive trajectories and replace a subset of historical ground-truth images with model-generated outputs. For selected rollout positions $\mathcal{S}_{\text{roll}}$, we define

$$\widetilde{I}_j = \begin{cases} \hat{I}_j, & j \in \mathcal{S}_{\text{roll}}, \\ I_j^*, & \text{otherwise}, \end{cases} \tag{10}$$

and train the subsequent image using

$$p_\theta(I_k^* \mid x, \widetilde{I}_{<k}). \tag{11}$$

The generated images are used only as historical conditions, while the supervision target remains the original ground-truth image. This exposes the model to realistic inference-time histories without directly treating its previous errors as targets. In practice, we retain persistent visual anchors and a small amount of clean early context while rolling out several later visual states, keeping the adaptation stage lightweight.

4 Experiments

4.1 Experimental Setup

Model and training. Our backbone is a 76B-parameter Mixture-of-Transformers (MoT) model with a 30B understanding branch (3B activated) and a 46B generation branch (7.6B activated). We natively train the model on diverse single-image and interleaved multimodal data, jointly learning multimodal understanding, autoregressive text generation, and high-fidelity image synthesis within a shared multimodal trajectory. The model supports multimodal contexts of up to 256K tokens.

Benchmark. We evaluate on **LongMIG-Bench**, which contains 100 stories evenly split between text-to-multi-image (T2MI) and image-to-multi-image (I2MI) generation, comprising 462 and 460 specified shots, respectively. Each story includes structured character, scene, and shot specifications, with recurring scenes designed to test visual consistency beyond a short recency window.

Baselines. We compare two groups of methods according to their evaluation protocol. For *Agent-Planned* generation, an external planner decomposes each story into per-image prompts, which are used to evaluate StoryDiffusion (Zhou et al., 2024), Qwen-Image (Zhao et al., 2026), UniSpace (Yan et al., 2026), HiDream (Cai et al., 2026), Lance-3B (Fu et al., 2026), and Show-o2 (Xie et al., 2025b). For native *Interleaved* generation, we evaluate BAGEL (Deng et al., 2025), SenseNova-U1.5 (Diao et al., 2026), ILLUME-X (Wang et al., 2026), Emu3.5 (Cui et al., 2025), and SEED-Story (Yang et al., 2025). We additionally reimplement UniLongGen (Chen et al., 2026a) on our backbone, denoted as UniLongGen*, for a controlled same-backbone comparison.

Evaluation. We use GPT-6 to evaluate **Character Identity**, **Scene Memory**, **Visual Fidelity**, **Visual Memory**, **Aesthetic Quality**, and **Text Rendering** on a 1–5 scale. Visual Memory specifically measures whether visual decisions introduced by earlier generated images persist later in the trajectory. We additionally report SceneGap using SigLIP2 (Tschannen et al., 2025) (**Gap-S**) and DINOv3 (Siméoni et al., 2025) (**Gap-D**), defined as the difference between average same-scene and cross-scene similarity.

Table 1: **Long-horizon generation on LongMIG-Bench.** Best and second-best results are shown in **bold** and underline, respectively.

Method	Character Identity ↑	Scene Memory			Visual Fidelity ↑	Visual Memory ↑	Aesthetic Quality ↑	Text Rendering ↑
		Scene ↑	Gap-S ↑	Gap-D ↑				
Agent-Planned								
StoryDiffusion	1.61	1.61	0.0615	0.1196	2.92	1.55	2.31	1.17
Qwen-Image	3.16	3.05	0.0694	0.2127	**3.83**	3.14	3.73	2.22
UniSpace	2.41	2.33	0.0346	0.1321	2.48	2.03	2.57	1.53
HiDream	3.35	2.58	0.0576	0.1620	3.19	2.76	2.71	2.78
Lance-3B	2.72	2.13	0.0480	0.1141	2.98	1.94	2.85	2.17
Show-o2-7B	2.72	2.90	0.0783	0.1966	2.89	2.37	2.83	1.00
Interleaved								
BAGEL-7B	3.38	1.44	-0.0099	-0.0191	2.84	1.70	2.52	1.35
SenseNova-U1.5	**3.72**	3.10	0.0062	0.0280	3.63	3.49	3.30	3.13
ILLUME-X	2.80	1.43	-0.0098	-0.0173	2.29	1.96	2.13	1.28
Emu3.5	3.51	2.97	0.0350	0.1077	3.73	3.24	3.69	2.15
SEED-Story	1.37	1.14	-0.0209	-0.0421	2.91	2.93	1.70	2.43
UniLongGen*	3.46	2.61	-0.0081	-0.0135	3.05	2.60	2.97	2.28
Ours	3.62	**3.35**	**0.0930**	**0.2499**	3.75	**3.77**	**3.75**	**3.50**

4.2 Main Results

Table 1 shows that our model combines strong long-range visual memory with high per-image generation quality. Ours achieves the best Scene score, Gap-S, Gap-D, Visual Memory, Aesthetic Quality, and Text Rendering among all evaluated systems, while ranking second in Character Identity and Visual Fidelity. In particular, the strong Visual Memory score indicates that the model preserves visual decisions introduced by earlier generated frames, rather than reconstructing each image independently from the textual specification. The same-backbone comparison with UniLongGen* further isolates the effect of visual-memory management. Under the same underlying model, our method improves Scene from 2.61 to 3.35, Visual Memory from 2.60 to 3.77, and Visual Fidelity from 3.05 to 3.75, while changing SceneGap from negative values to strongly positive separation under both SigLIP2 and DINOv3. This controlled comparison suggests that explicit memory formation and scene-aware retrieval provide benefits beyond training-free context curation.

Figure 3 further highlights the capability gap between simply supporting interleaved generation and sustaining high-quality generation over long trajectories. In image-conditioned generation, Ours preserves fine-grained facial appearance while producing diverse, photorealistic compositions rather than repeatedly copying a small set of poses or layouts. These results demonstrate that our setting goes beyond short or visually simplified interleaved generation and preserves strong native image-generation and multimodal-planning capabilities over extended trajectories.

4.3 Scaling Native Interleaved Generation to Longer Horizons

LongMIG-Bench focuses on a practical generation horizon that enables broad comparison across existing systems. We further examine whether the native interleaved capability of our model can persist substantially beyond this regime. As shown in Figure 4, the model can sustain autoregressive text-image generation over dozens of image-generation steps, including trajectories extending beyond 50 images. Across widely separated generations, recurring characters and visual states remain recognizable, while the model continues to introduce new compositions and content rather than simply reproducing earlier frames. These examples highlight an important property of our formulation: extending the generation horizon does not require truncating the multimodal trajectory. The complete semantic history remains accessible to the understanding pathway, while scene-aware visual memory controls the dense historical representations directly used for image synthesis.

4.4 Ablation Study

We conduct stage-specific controlled comparisons of scene-aware retrieval, persistent visual references, and self-rollout adaptation, as shown in Table 2. We use the 24-image Long+Ext stress test

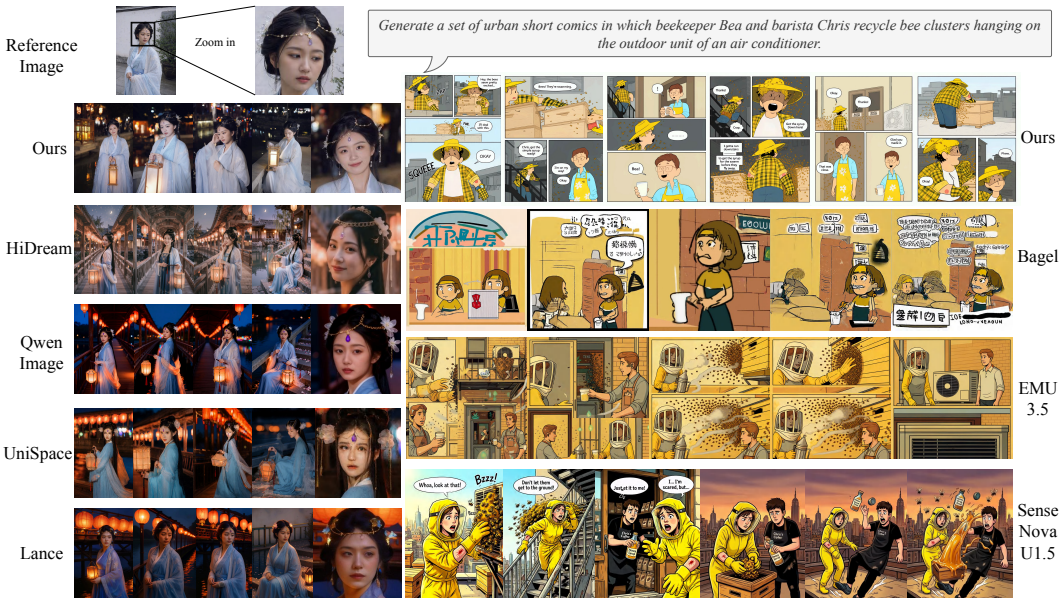


Figure 3: **High-fidelity long-form generation.** Left: image-conditioned generation with identity preservation across diverse views. Right: long-form interleaved generation with consistent characters and visual states over an extended trajectory.



Figure 4: **Native interleaved generation over extended trajectories.** Representative frames from long autoregressive generations across diverse visual domains, including 50+ image trajectories.

for these ablations, as the longer horizon more clearly exposes differences in memory behavior; LR metrics focus on scene revisits separated by at least eight generation steps.

Scene-aware retrieval. The first block evaluates how scene information affects visual-memory retrieval. Predicted routing performs comparably to retrieval using annotated scene labels, with small differences across scene and long-range metrics. In contrast, corrupting 30% of the scene signal reduces Scene from 3.26 to 2.81 and LR Scene from 3.10 to 2.76. Replacing scene-aware retrieval with fixed recent windows causes similar degradation, worsening when the window is reduced from four to two images. These results show that long-range generation benefits from retrieving semantically relevant historical states rather than relying on recent visual context.

Persistent visual references. The second block studies the role of persistent visual references. Removing reference-based training reduces Identity from 3.36 to 3.18, confirming that persistent references help stabilize recurring character appearance over long trajectories. Scene consistency also degrades, with Scene decreasing from 3.26 to 2.77 and LR Scene from 3.10 to 2.03; Gap-D and LR Gap-D similarly drop from 0.164 to 0.048 and from 0.119 to 0.005. In contrast, Visual Fidelity remains nearly unchanged (3.15 vs. 3.13). These results suggest that persistent references primarily provide stable identity anchors while also benefiting long-range visual consistency.

Table 2: **Controlled analysis of visual-memory components.** Ablations of scene-aware retrieval, persistent visual references, and self-rollout adaptation on the 24-image LONG+EXT stress test.

Variant	Setting	Identity ↑	Scene ↑	Gap-D ↑	LR Scene ↑	LR Gap-D ↑	Visual Fidelity ↑
Scene-aware retrieval							
Predicted Routing	Predicted Scene	**3.36**	3.26	0.164	3.10	0.119	**3.15**
GT Scene	Annotated Scene	3.21	**3.32**	**0.173**	**3.15**	**0.121**	3.03
Noisy Scene	30% Corrupted	3.30	2.81	0.145	2.76	0.099	3.14
Recent-4	Scene-Agnostic	3.15	2.85	0.147	2.79	0.092	2.97
Recent-2	Scene-Agnostic	3.05	2.58	0.114	2.47	0.071	2.89
Persistent visual references							
With Reference	Persistent Reference	**3.36**	**3.26**	**0.164**	**3.10**	**0.119**	**3.15**
w/o Reference	No Reference	3.18	2.77	0.048	2.03	0.005	3.13
Self-rollout adaptation							
Base	No Additional FT	2.89	2.57	0.193	2.51	0.157	2.57
GT-FT	GT History	3.23	2.71	0.191	2.52	0.161	3.35
Self-Rollout	Generated History	**3.59**	**3.25**	**0.194**	**3.12**	**0.174**	**3.38**



Figure 5: **Effect of self-rollout adaptation.** Base and GT-FT exhibit progressive reddish color drift and loss of fine-grained facial details, whereas Self-Rollout better preserves color appearance and local visual structure throughout the trajectory.

Self-rollout adaptation. The final block studies the effect of training on self-generated histories. Additional fine-tuning with ground-truth histories substantially improves Visual Fidelity over the base model (3.35 vs. 2.57), but provides only limited gains in long-range consistency. Under the same additional training data, Self-Rollout further improves Identity from 3.23 to 3.59, Scene from 2.71 to 3.25, and LR Scene from 2.52 to 3.12, while maintaining comparable Visual Fidelity (3.38). Gap-D and LR Gap-D remain relatively stable, indicating that the primary benefit of self-rollout lies in maintaining coherent identities and scenes when the model must continue from its own imperfect generations.

Figure 5 qualitatively illustrates this effect. Base and GT-FT exhibit color drift and loss of fine-grained facial details in later generations, whereas Self-Rollout better preserves color stability and visual structure. Together with the quantitative results, this suggests that exposure to self-generated histories is important for sustaining high-quality generation over long autoregressive trajectories.

5 CONCLUSION

We develop a large UMM natively trained for interleaved text-image generation and study how to extend its strong image synthesis and multimodal planning capabilities to long autoregressive trajectories. We introduce a visual-memory framework that combines persistent visual references, scene-aware retrieval, and self-rollout adaptation while preserving the complete semantic history for reasoning. Experiments on LongMIG-Bench demonstrate high-fidelity long-horizon generation with strong character and scene consistency, and controlled studies validate the complementary roles of the proposed components. Beyond the benchmark horizon, our model further sustains native interleaved generation over extended trajectories spanning dozens of images. Overall, our results show that explicit visual-memory management is important for scaling native interleaved generation to substantially longer horizons.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and formatting tasks. All research ideas, methodological decisions, experimental design, analyses, and conclusions were developed and verified by the authors. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

REFERENCES

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, et al. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Haoyu Chen, Qing Liu, Yuqian Zhou, He Zhang, Zhaowen Wang, Mengwei Ren, Jingjing Ren, Xiang Wang, Zhe Lin, and Lei Zhu. How long can unified multimodal models generate images reliably? taming long-horizon interleaved image generation via context curation. *arXiv preprint arXiv:2603.07540*, 2026a.

Leon Liangyu Chen, Haoyu Ma, Zhipeng Fan, Ziqi Huang, Animesh Sinha, Xiaoliang Dai, Jialiang Wang, Zecheng He, Jianwei Yang, Chunyuan Li, et al. Unit: Unified multimodal chain-of-thought test-time scaling. *arXiv preprint arXiv:2602.12279*, 2026b.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36:30286–30305, 2023.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025.

Jiaxing Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations*, volume 2026, pp. 85802–85822, 2026.

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.

Wenxun Dai, Zhiyuan Zhao, Yule Zhong, Yiji Cheng, Jianwei Zhang, Linqing Wang, Shiyi Zhang, Yunlong Lin, Runze He, Fellix Song, et al. Chatumm: Robust context tracking for conversational interleaved generation. *arXiv preprint arXiv:2602.06442*, 2026.

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.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, et al. Sensenova-u1.5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, et al. Lance: Unified multimodal modeling by multi-task synergy. *arXiv preprint arXiv:2605.18678*, 2026.

Lianyu Hu, Xiaoyu Ma, Zeqin Liao, and Yang Liu. Tvi-cot: Text-visual interleaved chain-of-thought reasoning for multimodal understanding. *arXiv preprint arXiv:2606.08464*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2025.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Luozheng Qin, GONG JIA, Yuqing Sun, Tianjiao Li, Haoyu Pan, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision. In *International Conference on Learning Representations*, volume 2026, pp. 15999–16028, 2026.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

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

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Chonghuinan Wang, Zhikai Chen, Chunwei Wang, Yecong Wan, Junwei Yang, Zhixin Wang, Wei Zhang, Jiaqi Xu, Renjing Pei, Xiaohe Wu, et al. Illuminating unified multimodal model for free-form interleaved text-image generation. *arXiv preprint arXiv:2606.30054*, 2026.

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*, 2024a.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37: 128374–128395, 2024b.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

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. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025a.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2025b.

Jinbo Xing, Zeyinzi Jiang, Yuxiang Tuo, Chaojie Mao, Xiaotang Gai, Xi Chen, Jingfeng Zhang, Yulin Pan, Zhen Han, Jie Xiao, et al. Wan-weaver: Interleaved multi-modal generation via decoupled training. *arXiv preprint arXiv:2603.25706*, 2026.

Jinbo Yan, Limeng Qiao, Jie Qin, Junyan He, Feize Wu, and Guanglu Wan. Unispace: Unified visual representation and scalable multimodal modeling. *arXiv preprint arXiv:2608.08676*, 2026.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1871–1881, October 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Mingcheng Ye, Jiaming Liu, and Yiren Song. Loom: Diffusion-transformer for interleaved generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4582–4592, June 2026.

Zijin Yin, Tiankai Hang, Yiji Cheng, Shiyi Zhang, Runze He, Yu Xu, Chunyu Wang, Bing Li, Zheng Chang, Kongming Liang, Qinglin Lu, and Zhanyu Ma. Generative visual chain-of-thought for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4657–4667, June 2026.

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, et al. Qwen-image-agent: Bridging the context gap in real-world image generation. *arXiv preprint arXiv:2606.26907*, 2026.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

A Additional Qualitative Results

We provide additional qualitative results to complement the quantitative evaluation in the main paper. We first compare complete long-form trajectories with representative baselines on multi-step visual process generation, then demonstrate generation across diverse visual domains, and finally present a complete native interleaved planning-and-generation trajectory produced by our model.

A.1 Long-Form Generation Comparisons

Figures 6 and 7 compare complete generation trajectories on two multi-step cooking tasks. Unlike isolated image generation, these examples require a model to represent distinct intermediate states and maintain meaningful progression throughout the sequence.

In the egg-tart example, several methods produce plausible individual images but repeatedly return to similar compositions or already completed food states, resulting in limited progression across the trajectory. Some interleaved models also reuse nearly identical visual layouts across multiple steps. Our model instead depicts a broader sequence of preparation, filling, baking, inspection, and serving stages, with substantial changes in viewpoint and composition across the trajectory.

The braised-pork example exhibits a similar distinction. The requested sequence requires the ingredients to undergo successive transformations rather than repeatedly showing the same intermediate or final dish. Our model progresses from ingredient preparation through cooking and thickening to the completed dish while maintaining a coherent illustrative style. Together, these examples show that strong individual images do not necessarily imply coherent multi-step visual progression over a long generation trajectory.

A.2 Generation Across Diverse Visual Domains

Figure 8 presents additional long-form trajectories across a diverse range of visual domains, including cinematic storytelling, cooking photography, minimalist typography, architectural visualization, illustrated recipes, graphic identity design, product design, and pet photography. These examples demonstrate that long-horizon interleaved generation is not limited to character-centric stories or a single visual style.

Different domains place different demands on long-form generation. Cinematic and photographic sequences require recurring subjects and environments to remain recognizable across changing viewpoints and compositions. Typography and branding tasks require consistent graphic language, layout, and text rendering across multiple outputs. Multi-step visual tasks require meaningful semantic progression, while image-conditioned product and pet examples require a user-provided visual identity to be preserved across substantially different contexts. The same unified model supports these different generation modes within a common interleaved framework.

A.3 Complete Multimodal Planning and Generation Trajectory

Figure 9 shows a complete native interleaved trajectory, including both intermediate textual planning and generated visual states. Given a request to create an eight-page nutrition presentation, the model first forms an overall content and visual plan, generates a reusable visual reference for the recurring character, and organizes the requested content into a sequence of slide-level scenes. It then alternates between textual planning and image generation until the complete presentation is produced.

The example illustrates how planning and visual synthesis occur within the same autoregressive multimodal trajectory. Each generated image is inserted back into the context and can influence subsequent planning and generation. Across the sequence, the model reuses the recurring character and overall graphic language while changing the slide content, layout, food imagery, and instructional message according to the evolving plan.

This differs from a pipeline in which an external planner first decomposes the task into independent prompts and a separate image generator renders each output in isolation. Here, intermediate text and generated visual states remain part of the same evolving multimodal context throughout generation.

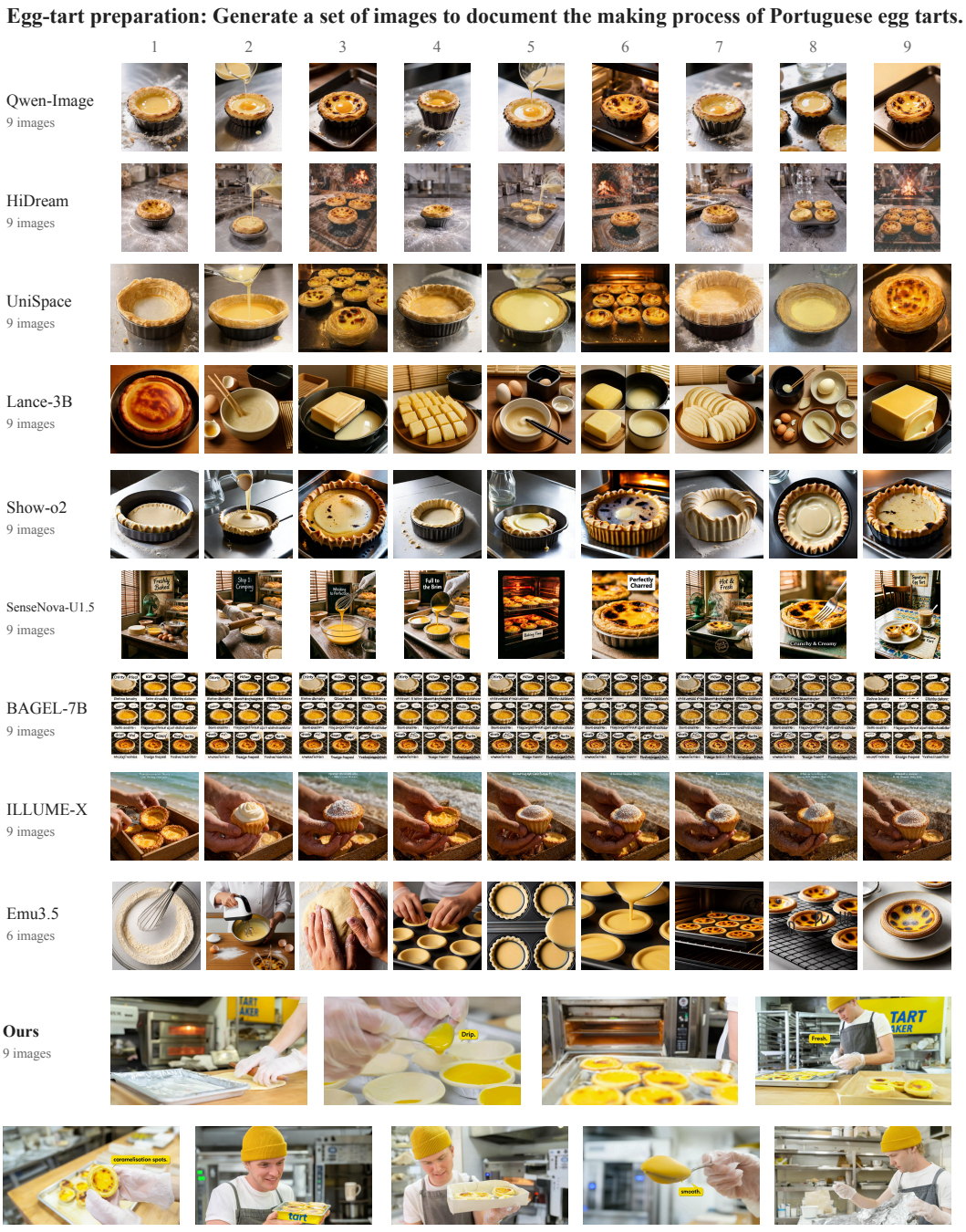


Figure 6: **Long-form visual process generation for Portuguese egg tarts.** All methods are prompted to generate a sequence documenting the preparation process. The comparison highlights differences in step progression, visual diversity, repetition, and consistency across the complete trajectory.

B Model and Training Details

B.1 Model Architecture

Our model is built on a Mixture-of-Transformers (MoT) architecture with separate modality-specific parameter streams for multimodal understanding and image generation. The understanding stream



Figure 7: **Long-form visual process generation for braised pork.** Each method generates a multi-image sequence describing the cooking process. The examples illustrate how models differ in representing successive visual states rather than repeatedly depicting similar intermediate or final results.

contains approximately 30B parameters with 3B activated parameters per token, while the generation stream contains approximately 46B parameters with 7.6B activated parameters. Both streams contain 48 Transformer layers and interact through joint attention over a shared multimodal sequence, while retaining separate projections, normalization layers, and feed-forward parameters.

As illustrated in Figure 10, the understanding stream processes text tokens and semantic visual tokens produced by a ViT encoder, whereas the generation stream processes dense VAE latents used for image synthesis. This separation directly supports the visual-memory formulation in the main paper: semantic visual representations remain accessible to the understanding pathway throughout the complete trajectory, while only the clean VAE representations of historical images are selectively exposed to the generation pathway.

Images are represented by two complementary encoders. A ViT encoder provides semantic representations for multimodal understanding, while a flow-matching VAE provides dense latent representations for image generation. At the 1K-resolution setting used in our experiments, an image contains approximately 4K VAE tokens and 1K semantic visual tokens. The model supports arbitrary aspect ratios under a fixed image-area budget and uses interleaved multimodal positional encoding for mixed text-image sequences.

Text generation is trained with an autoregressive language-modeling objective, while image generation uses a flow-matching objective. The combined objective is

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \lambda_{\text{img}} \mathcal{L}_{\text{img}}, \tag{12}$$

where $\lambda_{\text{img}} = 3$ in all experiments reported in this work.

B.2 Native Interleaved Training

The backbone is natively trained to process and generate mixed text-image sequences within a single autoregressive trajectory. Each training sample is represented as a multimodal conversation in which textual reasoning and image-generation calls are interleaved. Generated images are inserted back into the sequence and become available to subsequent multimodal understanding and image synthesis, allowing the model to learn sequential dependencies directly rather than treating each image as an independent generation problem.

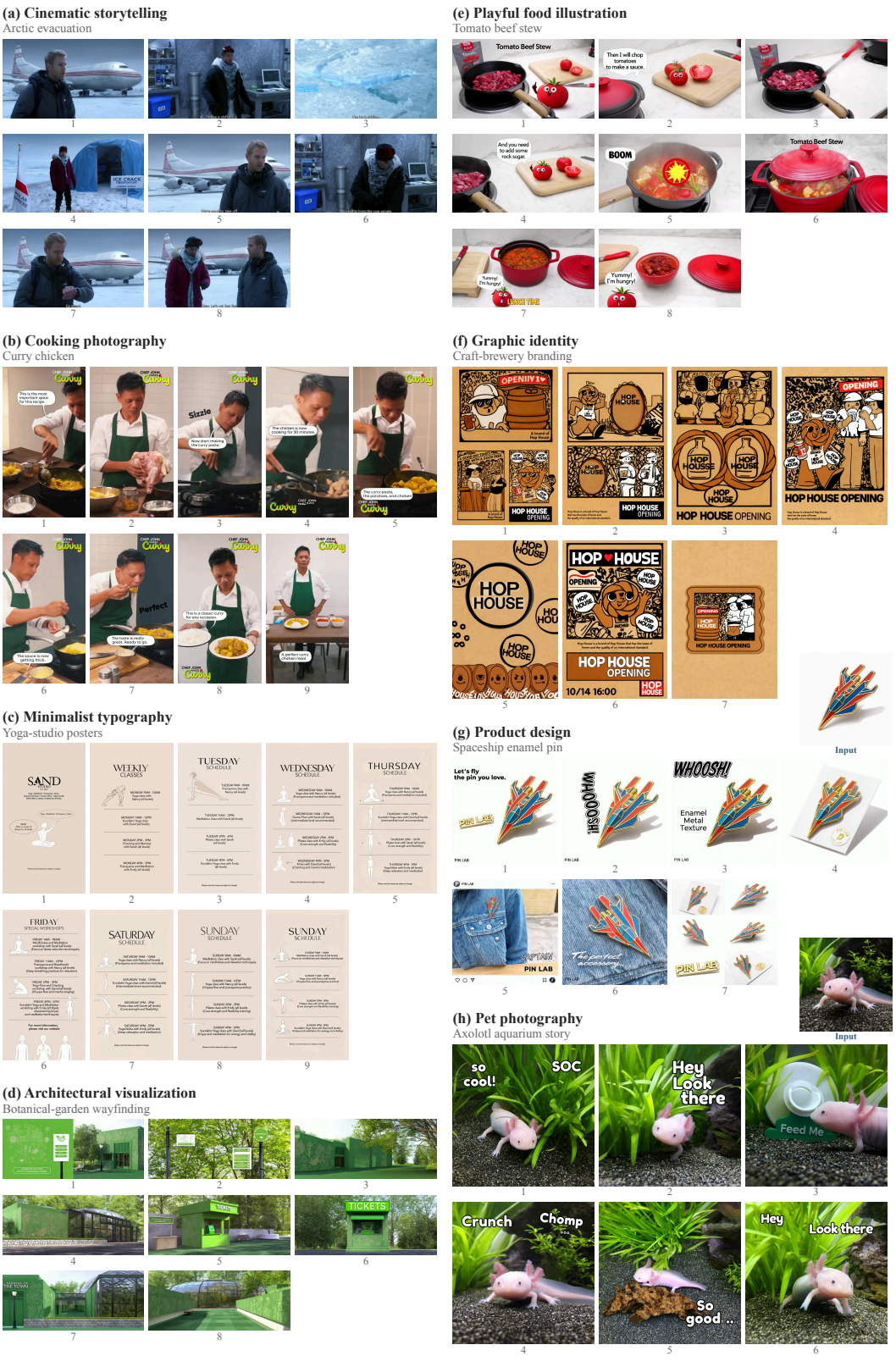


Figure 8: **Long-form generation across diverse visual domains.** Representative trajectories cover (a) cinematic storytelling, (b) cooking photography, (c) minimalist typography, (d) architectural visualization, (e) playful food illustration, (f) graphic identity design, (g) image-conditioned product design, and (h) image-conditioned pet photography. Input reference images for image-conditioned examples are shown separately.

Lunch nutrition and takeout choices: Create a 16:9 PowerPoint presentation for a nutrition lesson on lunch for office workers, 8 pages in total.

Bold: local CoT; regular: detailed descriptions.

000 Initial plan + reference 1

After receiving the user's request, I need to create a set of illustrated PPT content about lunch nutrition and takeout choices for office workers. This is a typical educational presentation, so I choose a 16:9 landscape widescreen format. It fits the conventions of PPT layouts and leaves ample breathing room when placing the main visual on the left and explanatory text on the right. To communicate the information vividly, I define a young white-collar boy representing office workers and name him Xiaoming. He appears through a combination of cartoon illustration and real photographs. His expressive gestures and speech bubbles make the dry topic of nutritional balance engaging and easy to understand.

The set contains eight images. The narrative begins by introducing the topic, then examines the components of a healthy lunch, provides practical advice, and finally presents a week of examples to put it into practice. For the frame design, all original images consistently use 16:9 landscape, which is well suited to PPT presentation. During character design, I generate an approximately 1:1 square reference for Xiaoming. The centered composition most clearly presents his delivery-worker uniform, facial details, and signature action of holding a phone, helping maintain character consistency in subsequent generation.

The following one character is designed:

- subject1 (name=Cartoon delivery boy): A young Asian male character is centered in a style combining cartoon illustration and real photography. He has short black hair and wears a dark-blue long-sleeved work uniform with a name badge on his chest and a dark-blue baseball cap. The front of the cap is embroidered with 'FRESH' in white, with a pale-blue circular logo above the brim. He holds a smartphone with a green border in his left hand to scan a code, while his right hand grips a white helmet with black 'FRESH' lettering on its side. His face is expressive, with wide-open eyes and an open mouth, and his right index finger points at his cheek, suggesting thought or surprise. He is placed against a pure white seamless background in a 1:1 square frame, focusing entirely on his action, clothing details, and the objects he holds.

Assistant: To complete this task, I need to use the generation tool to generate an image.

generate(source_image_indices_list=[], reception='', image_width=512, image_height=512, is_alpha=false)



Reference 1

002 Scenes + storyboard + panel 1

This story contains five scenes:

- Scene 1, Materials on a pure white background: A pure white background, with a top-down photograph of assorted ingredients on the left and explanatory text with a green background block on the right.
- Scene 2, Concepts on a pure white background: A pure white background, with a cartoon character illustration or a photograph with a yellow background block on the left and explanatory text on the right, without real food.
- Scene 3, At the office desk: A pure white background, with a desk carrying a yellow sticky note in the foreground and an open laptop on the desk.
- Scene 4, Takeout pitfalls on a pure white background: A pure white background, with photographs of various crossed-out unhealthy takeout foods on the left and explanatory text on the right.
- Scene 5, Weekly lunches on a pure white background: A pure white background, with photographs of various dishes on the left and explanatory text with a yellow background block on the right.

The story unfolds across eight images:

[1] (Reference: Xiaoming) Xiaoming sits at his computer looking at the screen, imagining what a perfect lunch should contain. He asks curiously, 'What makes a great lunch?'

[2] (Reference: -) A healthy lunch requires a carefully balanced combination. First come good staple foods such as brown rice, quinoa, or oats, which provide ample energy, just like the white rice in Xiaoming's hand.

[3] (Reference: -) Next come plentiful vegetables. Leafy greens and colorful peppers provide abundant vitamins and minerals, bringing vitality to the body.

[4] (Reference: -) Protein is also essential: chicken breast, fish, or tofu can help repair muscles. Finally, do not forget healthy fats such as avocado or olive oil to make the meal more complete.

[5] (Reference: Xiaoming) After learning the components, Xiaoming reflects on the takeout he usually orders. Holding his phone, he realizes with puzzlement that he often chooses burgers and fried chicken, thinking, 'What do I usually order?'

[6] (Reference: -) Xiaoming realizes that much takeout is unhealthy. Greasy fried chicken, high-calorie fries, and overly sweet doughnuts may be delicious, but eating them over a long period is bad for the body.

[7] (Reference: -) To avoid these pitfalls, Xiaoming decides to make a healthier plan. Looking at his computer screen, he lists foods he enjoys, such as sandwiches, salad, and curry, and decides to add them to his menu.

[8] (Reference: -) In this way, Xiaoming plans a week of lunches. From grilled chicken with brown rice on Monday to tuna salad on Tuesday and Thai curry on Wednesday, each meal is nutritionally balanced and gives him more energy for work!

[Scene 1: Materials on a pure white background]

In the first image, I need to establish the visual tone of the entire PPT set. This is a 16:9 landscape image with a left-right split composition. I place a colorful photograph of a food platter on the left and use a green background block behind the title on the right. Xiaoming appears in the lower-right corner looking at a computer screen, introducing the topic of what makes a great lunch through his dialogue.

This is a landscape PPT slide with a left-right split composition on a pure white background. The left shows a top-down view of a plate of food. A white round plate sits on a light-gray surface and contains several ingredients: brown-rice balls and yellow corn kernels at the bottom; orange-red grilled chicken pieces above, garnished with a small sprig of green parsley; halved avocados on the right, revealing yellow-green flesh and dark pits; and red tomato slices and a purple onion ring below. A silver cutlery cylinder holding a fork and spoon sits at the plate's right edge. Near the center-right is a green circular icon containing a smiling yellow sun. The right side mainly contains text. At the top is a three-line black English title: 'What makes', 'a great', and 'lunch?', with 'lunch?' backed by an irregular green brushstroke. Below are three smaller black lines: 'Let's explore what makes', 'a healthy lunch and', and 'how to order it.'. At the lower right, subject1 sits at a desk wearing a dark-blue uniform and a dark-blue baseball cap bearing 'FRESH' in white. His mouth is open and his eyes look at the laptop screen. His right index finger points toward the screen, and he holds a phone in his left hand. The laptop casing is silver-gray. To its left is a round speech bubble with a white fill and black border containing the black word 'Wow!'. The slide has

002 Scenes + storyboard + panel 1 (cont.)

bright, even lighting, a cheerful atmosphere, a colorful palette, and a left-right split composition.

Assistant: I will generate the image by calling the generation tool.

generate(source_image_indices_list=['epxk'], reception='the complete image description above (verbatim), image_width=1360, image_height=752, is_alpha=false)



Panel 1

004 Panel 2

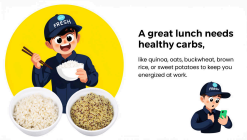
[Scene 2: Concepts on a pure white background]

In the second image, I begin breaking down the components of lunch. Still in 16:9 landscape, the left shows healthy grains and staple foods, while the right explains how they replenish energy. To enrich the visuals, I add a cartoon illustration against a yellow background block on the left, showing Xiaoming happily eating a bowl of rice and making the image more engaging.

This is a landscape PPT slide with a left-right split composition on a pure white background. The left contains a cartoon illustration and two food photographs. The illustration, set against a huge yellow circle, depicts subject1 with short black hair, large eyes, and an open mouth showing his teeth in a happy expression. He wears a dark-blue uniform and a dark-blue baseball cap with a pale-blue circular logo at its center containing 'FRESH' in white. His right hand holds a white bowl filled with white rice, and his left holds dark-brown chopsticks. Below him are two separate circular food photographs: brown rice in a white bowl on the left and quinoa in a white bowl on the right. All text on the right is black. At the top is a bold, large two-line heading: 'A great lunch needs' and 'healthy carbs,'. Below are three lines of smaller text: 'like quinoa, oats, buckwheat, brown', 'rice, or sweet potatoes to keep you', and 'energized at work.'. At the lower right is a smaller version of the same cartoon boy, wearing the same cap with 'FRESH' and holding a phone with a green casing. This slide is a non-scene design, so it has no shot scale or viewing angle. The overall lighting is bright and even, without distinct shadows. The atmosphere is lively and vivid, with a bright palette dominated by highly saturated yellow, blue, red, and green. The composition is a typical left-right column layout.

Assistant: I will call the generation tool to create the required image.

generate(source_image_indices_list=['epxk', 'pmal'], reception='the complete image description above (verbatim), image_width=1360, image_height=752, is_alpha=false)



Panel 2

006 Panel 3

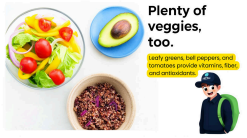
[Scene 2: Concepts on a pure white background]

The third image continues explaining the components of lunch, this time vegetables. Keeping the 16:9 frame, I show colorful photographs of vegetables on the left and text on the right. To echo the theme, I add an illustration of Xiaoming with a green backpack on the right, using his action to give concrete form to the concept of replenishing vitamins.

This landscape PPT slide or promotional poster uses a left-right column layout on a pure white background. The left is a photograph of food. A transparent glass bowl is filled with mixed vegetable salad, including red cherry tomatoes, green avocado pieces, yellow sweet-pepper strips, and white onion shreds. Beside it is a small blue round plate holding half an avocado with its brown pit exposed. In the foreground is a beige earthenware bowl containing dark-brown grain rice mixed with purple rice grains. The right contains text and a character illustration. At the upper right is a bold black three-line English heading: 'Plenty of', 'veggies,', and 'too.'. Beneath it is explanatory text on a yellow rounded rectangle, with three black lines: 'Leafy greens, bell peppers, and', 'tomatoes provide vitamins, fiber,', and 'and antioxidants.'. Below is subject1, with large eyes and a smile, wearing a dark-blue uniform and a dark-blue baseball cap. A pale-blue circular logo at the cap's center contains 'FRESH' in white. Subject1 wears a green backpack and faces toward the front left. The image has bright, even lighting and a fresh, healthy atmosphere. The overall palette is bright and the composition is split left and right.

Assistant: For this task, I need to call the generation tool to generate an image.

generate(source_image_indices_list=['epxk', 'pmal', 'phxe'], reception='the complete image description above (verbatim), image_width=1360, image_height=752, is_alpha=false)



Panel 3

008 Panel 4

[Scene 2: Concepts on a pure white background]

The fourth image completes the explanation of nutritional components. Within the 16:9 frame, the left shows meat and avocado, and the right contains explanatory text. To increase interaction, I place a waving Xiaoming on the right and add an English speech bubble with yellow highlighting beside him, making the explanation lively.

008 Panel 4 (cont.)

This is a landscape PPT slide with a left-right column layout on a pure white background. Three foods appear on the left: at the upper left, a white round plate holds skinned chicken pieces and chicken breast; near the middle left is a dark-gray brown-rice ball; and at the lower left, a small blue bowl holds sliced avocado sprinkled with black and white sesame seeds. The right contains text and a character illustration. At the upper right are three bold black English heading lines: 'Lean protein', 'for muscle', and 'repair,'. Below is a yellow rounded rectangle containing three smaller black lines: 'like chicken, fish, beans, tofu, or', 'eggs, paired with healthy fats from', and 'avocados, nuts, or olive oil.'. At the lower right, subject1 smiles and raises his right hand in greeting. A circular logo at the center of his cap contains 'FRESH' in white. An oval speech bubble extends from subject1, containing three lines of black text in which 'healthy,' is highlighted in yellow. The full text reads 'A great lunch includes healthy carbs, plenty of veggies, lean protein for muscle repair, and healthy fats.', with the word 'healthy,' emphasized by a yellow background. This PPT slide is a non-scene design, with bright, even lighting and a healthy, positive atmosphere. The overall palette is bright and the composition uses columns.

Assistant: To generate this image, I will call the generation tool.

generate(source_image_indices_list=['epxk', 'pmal', 'phxe', 'erkt'], reception='the complete image description above (verbatim), image_width=1360, image_height=752, is_alpha=false)



Panel 4

010 Panel 5

[Scene 3: At the office desk]

In the fifth image, the scene shifts to the desk for a moment of self-reflection. In the 16:9 landscape composition, the left shows a half-length view of Xiaoming thinking while looking at his computer screen. A yellow sticky note in the foreground adds an office atmosphere. Photographs of a burger and fried chicken on the right represent his usual unhealthy lunch choices, raising the question, 'What do I usually order?'

This is a landscape PPT slide with a left-right column layout on a pure white background. The left depicts subject1 with short brown hair, fair skin, thick dark eyebrows, wide eyes, and a slightly open mouth, looking intently at the computer screen in front of him. He wears a dark-blue long-sleeved uniform and a matching baseball cap. At the center of the cap is a pale-blue circular logo containing 'FRESH' in white. His right hand is raised, with his index and middle fingers lightly touching his chin and lips, while his left rests on the desk in front of him. He sits at a white desk with an open silver-gray laptop; the keyboard is not visible. The lit screen shows a colorful sunset seascape with an orange-yellow sky, clouds, and shimmering sea. A black question mark, '?', appears above and to the left of his head. On the desk to his front right is a square yellow sticky note with 'Lunch' in bold black English text. The right side shows two fast foods. On the left is a double cheeseburger with a golden bun, two thick beef patties, melted yellow cheese, red ketchup, and green lettuce. A pink carton with white polka dots sits beneath it. On the right are fried chicken pieces in a yellow wavy-edged fries carton. A red label at the top reads 'FRESH' in white. Between the burger and chicken, near the center-right, two lines of black text read 'What do I' and 'usually order?', with 'usually' highlighted by an irregular green painted shape. This slide is a non-scene design, so it has no shot scale or viewing angle. Lighting is even and bright, without distinct shadows. The image conveys curiosity. Its subjects have vivid colors and a highly saturated palette. The composition is a typical left-right column layout.

Assistant: I will continue generating images by calling the generation tool.

generate(source_image_indices_list=['epxk', 'pmal', 'phxe', 'erkt', 'cvmd'], reception='the complete image description above (verbatim), image_width=1360, image_height=752, is_alpha=false)



Panel 5

012 Panel 6

[Scene 4: Takeout pitfalls on a white background]

The sixth image introduces pitfalls to avoid. In 16:9 landscape, the left shows unhealthy foods such as fries and doughnuts crossed out in red, directly visualizing the concept of avoiding pitfalls. The right contains explanatory text. The overall palette is bright but carries a warning, reminding viewers to choose carefully.

This is a landscape promotional poster showing several prohibited foods on a white background. Two lines of black text in the upper-left area read 'Not all takeout is healthy.' and 'Here's what to avoid:'. Four foods appear near the center-left. At the top is a portion of golden fries in a carton with blue edging and red polka dots. A red label on the front reads 'FRESH' in white. A large red cross appears above the fries. To their right is a box of fried chicken pieces in a yellow carton with a wavy edge. A red oval label on its front reads 'FRESH' in white. A red cross also appears above the chicken. At the lower left is a pink doughnut covered in thick pink icing and colorful sprinkles, with a red cross below it. At the lower right is a burger in a transparent plastic bowl, containing bread, lettuce, tomato slices, a meat patty, and melted cheese. A red cross also appears below the burger. At the middle right is a cartoon subject1 with large eyes and a smile. He wears a dark-blue long-sleeved top and a dark-blue baseball cap. A blue-green circular logo at the center of the cap contains the white word 'FRESH'. He raises both hands in front of his chest in a gesture of refusal. A speech bubble beside his right arm points toward him; its two lines read 'Skip the' and 'fries!'. The poster uses a diagonal composition, bright and even lighting, and vivid, lively colors to create a lively atmosphere.

Assistant: To handle this task, I need to call the generation tool to create an image.

012 Panel 6 (cont.)

generate(source_image_indices_list=['epxk', 'pmal', 'phxe', 'erkt', 'cvmd', 'kwwh'], reception='the complete image description above (verbatim), image_width=1360, image_height=752, is_alpha=false)



Panel 6

014 Panel 7

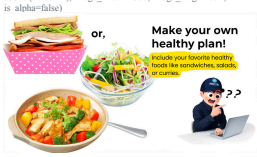
[Scene 2: Concepts on a pure white background]

The seventh image proposes a solution. In 16:9 landscape, the left shows photographs of healthy foods such as sandwiches, salad, and curry, representing items Xiaoming wants to add to his menu. The right contains text and an illustration of Xiaoming thinking at his computer. His thoughtful action guides viewers to develop their own healthy lunch plan.

This infographic promotes a healthy lunch plan using a left-right column layout on a pure white background. Three foods appear on the left. At the upper left is a double-layer sandwich in a pink carton with white polka dots. Two slices of bread enclose lettuce, sliced meat, cheese, tomato, and sauce. Black text reading 'or,' appears at the upper right. In the middle is vegetable salad in a transparent glass bowl, containing green, yellow, and red pepper strips, purple onion shreds, white bean sprouts, and leafy greens. Below is Thai curry in a beige ceramic bowl, containing chicken pieces, diced carrots, broccoli, and potato chunks, topped with cilantro and accompanied by a silver spoon. The right side contains text and a character. A large bold black heading reads 'Make your own healthy plan!'. Below is an irregular yellow highlighted area containing three lines of small black text: 'Include your favorite healthy foods like sandwiches, salads, or curries.'. Beneath the text is subject1 with large eyes, a smile, and his right index finger pointing at his cheek. He wears a dark-blue baseball cap with a circular logo on the front containing 'FRESH' in white. He wears a dark-blue long-sleeved shirt and sits at a computer with both hands on the keyboard. Two black question marks, '??', appear above his head. This infographic is a non-scene design with bright, even lighting, a cheerful atmosphere, a colorful palette, and a left-right column composition.

Assistant: I will use the generation tool to create the image as requested.

generate(source_image_indices_list=['epxk', 'pmal', 'phxe', 'erkt', 'cvmd', 'kwwh', 'ezks'], reception='the complete image description above (verbatim), image_width=1360, image_height=752, is_alpha=false)



Panel 7

016 Panel 8

[Scene 5: Weekly lunches on a white background]

The eighth image summarizes and presents the result. In 16:9 landscape, the left shows photographs of five different healthy lunches, while the right lists the week's menu. A grid layout makes the information clear and easy to follow, bringing this nutrition PPT set to a satisfying close.

This infographic presents a week's healthy lunch menu using a grid layout on a pure white background. Five meals are arranged from top to bottom on the left: at the upper left, a white round plate holds grilled chicken, brown rice, avocado slices, corn kernels, onion rings, and tomato slices; near the middle left, a black plate holds a sandwich, potato wedges, and carrot sticks; at the upper right, a white round plate holds tuna salad, brown rice, and broccoli; below, a white round plate holds curry chicken-leg rice, brown rice, and green peas; at the bottom, a white round plate holds baked fish, brown rice, avocado, and corn. The right side contains explanatory text. At the top is a bold black heading: 'Here's my weekly lunch plan:'. Below is a table listing lunch arrangements from Monday to Sunday, with a day, food name, and description in each row. The specific content is as follows: first row, 'Monday: Grilled chicken with brown rice and veggies.' Second row, 'Tuesday: Tuna salad with quinoa.' Third row, 'Wednesday: Thai curry with brown rice.' Fourth row, 'Thursday: Turkey sandwich with salad.' Fifth row, 'Friday: Baked fish with brown rice and avocado.' To the right of the table is a vertical column of black text: 'A healthy lunch fuels your afternoon energy and focus!'. This is an infographic rather than a depicted scene. Lighting is bright and even, the atmosphere is healthy and orderly, and the overall palette is bright.

Assistant: I need to call the generation tool to generate this image.

generate(source_image_indices_list=['epxk', 'pmal', 'phxe', 'erkt', 'cvmd', 'kwwh', 'ezks', 'uakf'], reception='the complete image description above (verbatim), image_width=1360, image_height=752, is_alpha=false)



Panel 8

018 Completion

I have completed image generation and can reply to the user.

Assistant: Image creation is complete; here are the results.

Figure 9: **Complete native interleaved planning-and-generation trajectory.** Given a request for an eight-page nutrition presentation, the model first forms an overall plan and generates a reusable visual reference, then alternates between textual planning and image generation until completing the requested sequence. Generated images are inserted back into the multimodal context and remain available to subsequent generation steps.

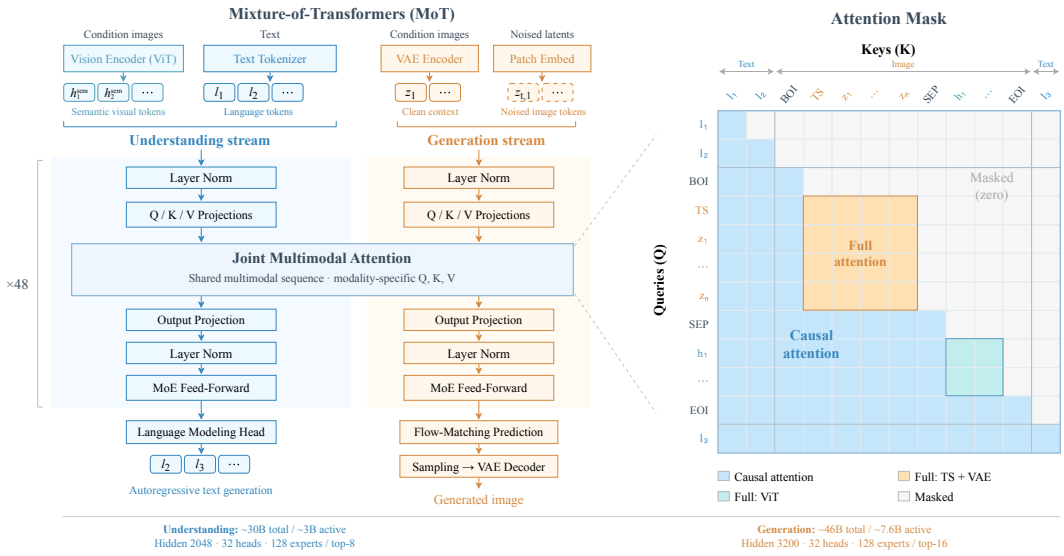


Figure 10: **Architecture and attention pattern of the MoT backbone.** The understanding stream processes language and semantic ViT tokens, while the generation stream processes clean conditioning latents and noised image latents. The two streams interact through joint multimodal attention with modality-specific projections and feed-forward layers. Right: the mixed attention pattern used for interleaved text-image modeling.

Table 3: **Architecture of the two MoT streams.** Activated parameters are reported per token.

	Understanding	Generation
Layers	48	48
Hidden size	2048	3200
Attention heads	32	32
Routed experts	128	128
Experts per token	8	16
Total parameters	∼30B	∼46B
Activated parameters	∼3B	∼7.6B

Each generated image has two roles in the training sequence. A noised latent is used as the flow-matching target for the current generation step, while a clean copy is retained as visual context for subsequent generations. Semantic ViT representations of all historical images remain visible to the understanding stream, whereas the clean VAE representations form the generative visual states controlled by our memory mechanism.

Two annotations are associated with long-form training trajectories. Character-reference images are marked as persistent visual references and remain accessible throughout generation. Each story image is additionally assigned a discrete scene identifier indicating its physical scene, which supervises scene-state prediction during training and enables scene-aware retrieval at inference time.

B.3 Long-Horizon Post-Training

For long-horizon adaptation, we post-train the model exclusively on interleaved story trajectories containing recurring entities, scene transitions, and long-range revisits. The training sequence jointly supervises textual planning, visual-reference formation, scene-state prediction, and image generation. Scene-aware visual retrieval uses a compact fixed budget that combines a short recency window with historical states associated with the predicted scene. Specifically, the model retains recent visual context while retrieving both early establishing states and more recent states from the same scene, allowing it to recover long-range visual information without exposing the full generative history. Persistent visual references remain visible independently of this retrieval budget. Training

sequences are packed to a maximum length of approximately 256K tokens, with cross-story attention blocked and positional indices reset for each story.

Optimization. We use a hybrid Muon–AdamW optimization setup. Muon uses a momentum of 0.95, while AdamW uses $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and a weight decay of 0.1. Gradients are clipped to a maximum norm of 1.0. Long-horizon post-training uses a constant learning rate of 3×10^{-5} after a 200-step warmup.

B.4 Self-Rollout Training

We perform self-rollout as a lightweight few-step post-training stage on top of the teacher-forced model. Here, *few-step* refers to the small number of additional optimization steps required for adaptation, rather than a shortened rollout horizon or fewer image-generation steps at inference time. In practice, the model typically converges within roughly 500 post-training steps.

Model-generated images are first produced offline and written back into the training trajectories as alternative historical conditions. During training, only the clean conditioning copy of a selected historical image is replaced by the model-generated result; the noised image used as the supervision target always remains ground truth. The model is therefore trained to recover the correct target from imperfect generated history without directly distilling its own previous errors.

Rollout substitution is applied probabilistically, while persistent visual references and the first few story images are kept clean to preserve reliable identity and scene initialization. The understanding stream and semantic visual encoder are frozen during this stage, and optimization focuses on adapting the generation pathway to imperfect autoregressive histories. Our ground-truth fine-tuning (GT-FT) control uses the same additional training data and optimization budget as self-rollout, but retains ground-truth images as historical conditions throughout training. This isolates the effect of replacing clean teacher-forced histories with self-generated conditioning, rather than the effect of additional fine-tuning itself.

C Training Data Construction

C.1 Data Pipeline and Scale

The overall training pipeline uses progressively curated multimodal data, moving from broad multimodal coverage toward increasingly specialized long-horizon supervision. We begin with a large-scale candidate pool on the order of 10^8 examples, covering natural images, image–text pairs, image-conditioned generation data, multi-image sequences, comics, illustrated stories, and design-oriented content. The candidate pool is progressively processed through quality filtering, deduplication, resolution and file-integrity checks, semantic consistency filtering, and task-specific parsing.

The resulting data are organized into several stages with increasingly specialized supervision. Pre-training (PT) uses data on the order of 10^6 examples to learn general multimodal understanding and generation. Supervised fine-tuning (SFT) further uses on the order of 10^5 curated examples spanning single-image generation, image-conditioned generation, multi-image generation, and native text–image interleaving. From this mixture, we construct a dedicated long-horizon corpus containing recurring entities, explicit scene transitions, and long-range scene revisits. Finally, self-rollout adaptation uses a smaller set of trajectories augmented with the model's own generated histories.

C.2 Long-Horizon Corpus Composition

The long-horizon post-training corpus contains $21{,}162$ interleaved stories with $197{,}653$ generated images, including $53{,}499$ persistent visual reference images and $144{,}154$ story panels. Reference images precede the corresponding story panels and provide reusable visual states for recurring entities. Each story contains between 2 and 24 story panels and between 1 and 10 visual references, with a mode of two.

The corpus is constructed from several types of sequential visual data, including multi-shot storyboards, illustrated children's books, social-media comics, long-form comics with recurring environments, and design-oriented sequences. The source distribution is summarized in Table 4.

Table 4: **Composition of the long-horizon post-training corpus.**

Data type	Fraction
Multi-shot storyboards	45.3%
Illustrated books with visual references	18.5%
Illustrated books without visual references	7.9%
Social-media comics	14.5%
Long comics with background continuity	9.1%
Brand and design sequences	4.7%

For sequential sources, we additionally parse recurring entities and physical scenes and convert the retained examples into a unified long-form interleaved format. Samples with missing images, ambiguous or unresolvable scene structure, duplicate content, or trajectories exceeding the training budget are removed during this process.

C.3 Interleaved Sequence Construction

Each long-horizon example is represented as a multi-turn multimodal conversation in which textual planning and image generation are interleaved within a single trajectory. Before generating an image, the assistant may produce textual reasoning describing the intended content and its relation to the preceding trajectory, followed by an image-generation call. The generated image is then inserted back into the sequence and becomes available to subsequent text and image generation.

A trajectory can be written schematically as

$$x \to r_1, I_1 \to r_2, I_2 \to \cdots \to r_N, I_N, \tag{13}$$

where x is the initial instruction, r_k denotes the textual planning associated with generation step k, and I_k is the corresponding generated image.

Every generated image participates in the sequence in two complementary forms. A noised latent is used as the flow-matching target for the current generation step, while a clean visual representation is retained as historical conditioning for subsequent generations. Thus, later generations can directly depend on visual states produced earlier in the same trajectory rather than treating each image as an independent training example.

Later image-generation calls also retain references to previously generated images, allowing the training sequence to preserve the complete multimodal trajectory and to supervise dependencies spanning multiple generation steps.

C.4 Persistent References and Scene Annotations

Long-form trajectories contain two additional types of structure used by the visual-memory mechanism.

First, auxiliary visual-reference images are identified as *persistent visual references*. These images are generated before the corresponding story panels when recurring visual entities need to be established and are trained with the same native image-generation objective as ordinary outputs. Unlike story panels, they are treated as persistent generative memory and remain accessible throughout the subsequent trajectory.

Second, each story panel is assigned a discrete scene state according to its underlying physical environment. Images depicting the same physical location retain the same scene state despite changes in camera viewpoint, composition, character pose, or local action. Persistent reference images are excluded from scene matching and are handled independently by the persistent-memory rule.

These annotations are used to supervise scene-state prediction during long-horizon post-training. At inference time, no ground-truth scene annotation is provided: the model predicts the upcoming scene state from its multimodal history, and this prediction is subsequently used for scene-aware visual-memory retrieval.

Table 5: **Composition of LongMIG-Bench.**

	T2MI	I2MI
Stories	50	50
Specified shots	462	460
Avg. shots / story	9.2	
Min. revisited shots / story	≥ 4	
Avg. revisited shots	6.16	6.04
Distinct styles	50	50
User reference image	No	Yes

For the no-reference ablation, we rebuild the same training episodes after removing all persistent-reference generation calls while preserving the remaining story panels and their ordering.

C.5 Self-Rollout Data

Self-rollout uses a parallel version of the long-horizon corpus in which selected historical conditioning images are replaced by outputs generated by the model itself. The rollout images are produced offline and written back into the corresponding training trajectories; no autoregressive image sampling is performed inside the optimization loop.

The two rollout stages contain $5{,}721$ and $7{,}522$ stories, respectively, for a total of $13{,}243$ rollout trajectories. The second stage places greater emphasis on portrait and social-media sequences, where identity drift and visual error accumulation are particularly common.

For a selected rollout position, only the clean historical conditioning copy is replaced by the model-generated image. The noised image used as the flow-matching supervision target remains the original ground-truth image. Persistent visual references are never replaced, and the first two non-reference story images are kept clean to provide a reliable initial visual state. This construction exposes the model to realistic inference-time histories without directly treating its own previous errors as training targets.

D LongMIG-Bench Construction

D.1 Benchmark Composition

LongMIG-Bench contains 100 long-form story specifications, evenly divided between text-to-multi-image (T2MI) and image-to-multi-image (I2MI) generation. The T2MI track contains 50 stories with 462 specified shots, while the I2MI track contains 50 stories with 460 shots and additionally provides a user reference photograph for each story. Stories contain 9.2 shots on average and cover illustrated children's books, comics, storyboards, portrait series, pet and product photography, recipes, poster design, sticker sets, and landscape compositions.

To avoid evaluating only a narrow visual regime, each story specifies its own visual style and target aspect ratio. The benchmark contains 50 distinct styles in each track and seven aspect ratios ranging from $9{:}16$ to $16{:}9$. Every specification is available in both English and Chinese.

D.2 Structured Story Specifications

Each story is represented as structured data rather than free-form prose. A specification contains a cast list with per-character visual attributes, a location table, an ordered shot sequence, and shot-level annotations describing the corresponding scene, whether the scene revisits an earlier location, and which entities should persist from previous shots. This structure provides explicit supervision for evaluating identity preservation, scene recurrence, and long-range visual consistency.

A simplified specification can be represented as

$$\mathcal{S} = \{\mathcal{C}, \mathcal{L}, (q_1, s_1, r_1, e_1), \ldots, (q_N, s_N, r_N, e_N)\}, \tag{14}$$

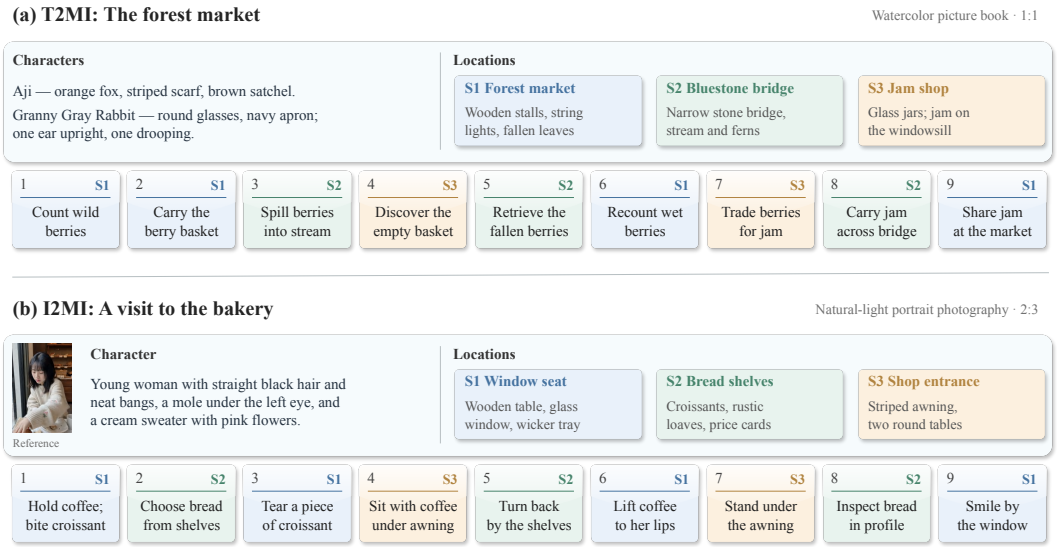


Figure 11: **Examples from LongMIG-Bench.** Each story contains structured character and location definitions together with an ordered sequence of shots. Colors indicate recurring scenes, and arrows highlight long-range revisits after intervening scene transitions. The I2MI track additionally provides a user reference image.

where $\mathcal{C}$ and $\mathcal{L}$ denote the character and location definitions, q_k is the specification for shot k, s_k is its scene identifier, r_k indicates whether the scene is a revisit, and e_k specifies entities carried over from the preceding trajectory.

D.3 Long-Range Revisit Design

Long-range scene revisits are introduced deliberately rather than occurring incidentally. Every story contains at least four revisited shots, with an average of 6.16 revisits in T2MI and 6.04 in I2MI. In these trajectories, scenes established earlier in the story reappear after multiple intervening shots and scene transitions.

This design is important for evaluating visual memory. If scenes occur only once, or if every revisit falls within a short recency window, scene-aware retrieval becomes difficult to distinguish from simply retaining the most recent images. LongMIG-Bench instead contains repeated long-range revisits for which the relevant visual state may lie far outside the recent context, directly testing whether a model can recover previously established visual information.

D.4 Benchmark Examples

Figure 11 visualizes representative specifications from both benchmark tracks. In addition to the natural-language story instruction, we show the underlying character definitions, scene structure, shot sequence, and long-range revisit relationships. These structured annotations are used only for benchmark construction and evaluation; models receive the corresponding story-level generation instruction rather than ground-truth scene annotations during open-ended generation.

D.5 Long+Ext Long-Horizon Stress Test

In addition to LongMIG-Bench, we construct Long+Ext, a locked-text stress-test suite for controlled long-horizon evaluation. Every Long+Ext story contains exactly 24 predefined image slots with fixed shot descriptions and scene assignments.

Long+Ext contains 38 stories in total. Each story contains two recurring characters and three recurring scenes that are repeatedly revisited throughout the 24-image trajectory. The scene-transition

patterns are fixed during benchmark construction to ensure frequent non-adjacent revisits rather than contiguous scene blocks.

All methods are evaluated on the same shot sequence and scene structure. Native interleaved models receive the complete story-level instruction, whereas independent image generators receive corresponding self-contained per-shot prompts.

Long-range (LR) metrics evaluate only revisits occurring at least eight generation steps after the *first establishing view* of the corresponding scene:

$$k - e_s \geq 8, \tag{15}$$

where e_s is the first occurrence of scene s. This setting specifically tests whether a model can recover distant visual states beyond a short recent-context window. LONG+EXT is used primarily for the controlled component ablations in Table 2.

E EVALUATION METRICS

Character Identity. **Character Identity** measures whether each named recurring character remains visually recognizable across the frames in which it appears. The evaluator considers identity-defining attributes such as facial appearance, body shape, hair or fur characteristics, clothing, and other persistent details, while allowing pose, expression, viewpoint, and illumination to vary. Repeating nearly identical images can trivially inflate identity consistency, so this metric is interpreted jointly with Scene Memory and Visual Memory.

Scene Memory. **Scene Memory** measures whether a previously established physical environment is correctly recovered after the trajectory leaves and later revisits it. The evaluator considers persistent architectural structure, spatial layout, color palette, signage, and other location-specific visual cues, while also requiring distinct environments to remain distinguishable. To prevent degenerate background copying, a trajectory that effectively reuses the same background for all scenes receives a Scene score of at most 2.

For long-range analysis on LONG+EXT, **LR Scene** restricts evaluation to revisits occurring at least eight generation steps after the first establishing view of the corresponding scene.

Visual Memory. **Visual Memory** measures whether later generations preserve visual decisions introduced by earlier generated frames rather than reconstructing each image solely from the textual specification. Such information may include the position of a facial mark, the layout of a storefront sign, a garment pattern, or other instance-specific details that emerge during generation. Preserving identity directly from a user-provided reference image does not by itself constitute visual memory. Conversely, trivially repeating the same frame throughout the trajectory is assigned the minimum score.

Visual Fidelity and Aesthetic Quality. **Visual Fidelity** evaluates the technical quality of generated images within the visual style requested by the story, including anatomy, composition, rendering artifacts, and malformed text. Stylized, grayscale, comic, watercolor, or other non-photorealistic outputs are evaluated relative to their requested style rather than against a photorealistic standard.

Aesthetic Quality evaluates overall visual craftsmanship and appeal within the declared style, including composition, graphic coherence, visual polish, and style-appropriate presentation. Photorealism is not rewarded unless explicitly requested. Severe noise, malformed anatomy, corrupted text, inappropriate borders, or repetitive use of a single visual layout reduce the score.

Text Rendering. **Text Rendering** evaluates the visual quality, legibility, and semantic faithfulness of story-relevant text rendered within generated images, including storefront signs, posters, menus, titles, dialogue, and other scene-integrated textual elements. The evaluator considers whether characters and words are clearly formed, readable, visually coherent with the surrounding content, and free from common rendering artifacts such as malformed glyphs, distorted strokes, or inconsistent typography. This metric reflects the model's ability to preserve strong text-generation capability within long-horizon interleaved image sequences.

SceneGap. We additionally measure scene separability using

$$\mathrm{SceneGap} = \bar{s}_{\mathrm{same}} - \bar{s}_{\mathrm{cross}}, \tag{16}$$

where $\bar{s}_{\mathrm{same}}$ is the average visual similarity between images associated with the same scene and $\bar{s}_{\mathrm{cross}}$ is the average similarity between images from different scenes. We compute SceneGap using SigLIP2 and DINOv3 representations, denoted **Gap-S** and **Gap-D**, respectively. Positive values indicate that recurring scenes are visually closer to themselves than to other environments.

E.1 Trajectory-Level Judge

We use GPT-6 as the trajectory-level evaluator for Character Identity, Scene Memory, Visual Memory, Visual Fidelity, Aesthetic Quality, and Text Rendering. The evaluator receives the complete story specification together with the generated trajectory and assigns an absolute score from 1 to 5 for each applicable dimension. Scores are assigned at the trajectory level rather than by enforcing a potentially incorrect one-to-one alignment between generated images and specified shots.

The evaluation rubric explicitly distinguishes consistency from degenerate repetition. Reusing nearly identical images may preserve identity but does not demonstrate visual memory or scene differentiation. Similarly, a sequence that reuses essentially the same environment throughout the story is capped at 2 for Scene Memory. Visual Fidelity and Aesthetic Quality are judged relative to the visual style requested by the specification, so intentionally stylized or non-photorealistic outputs are not penalized for deviating from photographic appearance.

E.2 Evaluator Prompts

For reproducibility, we provide the complete prompts used by the GPT-6 trajectory-level evaluators. Generated frames are supplied together with the textual prompt, and the evaluator is required to return structured JSON. We use three specialized evaluators: the primary trajectory evaluator for Character Identity, Scene Memory, and Visual Fidelity; a second evaluator for Visual Memory and Aesthetic Quality; and a third evaluator dedicated to Text Rendering.

Primary trajectory evaluator: Identity, Scene, and Visual Fidelity

```
You are an expert evaluator of AI-generated visual stories (storyboards, picture books,
     comics).

You will be given (1) a STORY SPEC that every system was asked to render, and (2) the
     frames ONE system produced, in the order it produced them.

CRITICAL FRAMING RULES
- The number of frames may differ from the number of shots in the spec. These systems plan
      their own storyboards. Do NOT reward or penalise a system merely for frame count,
     and do NOT assume frame k corresponds to shot k. Work out the correspondence yourself
      from the content.
- Judge the frames as a SEQUENCE, not as isolated pictures.
- Judge craft against the story's DECLARED art style. If the spec asks for flat colour,
     watercolour, greyscale or black-and-white manga, then correctly following that is
     GOOD. Never reward photorealism or saturated colour for its own sake.
- Ignore any watermark, site header or page furniture.
- Do not give courtesy scores. Use the full 1-5 range. 3 means "acceptable but with clear
     defects".

Score each axis as an INTEGER 1-5.

1. character_identity
Is each named cast member the SAME individual in every frame where they appear (face/
     muzzle structure, body proportions, hair or fur colour, clothing and its details,
     accessories)? Pose, expression, camera distance and lighting are ALLOWED to change
     and must not be penalised.
5 = a reader would never doubt it is the same character.
4 = one or two small slips, e.g. an accessory changes.
3 = recognisable but clearly drifting.
2 = frequently reads as a different character.
1 = essentially a new character each frame, or the species/kind changes.

2. scene_memory
When the story RETURNS to a location after visiting other locations, does that location
     read as the SAME place (architecture, layout, palette, signage, landmark props)? AND
     do DIFFERENT locations stay clearly DISTINCT from each other?
```

```
5 = revisits are unmistakably the same place AND distinct locations are clearly different.
3 = revisits are only roughly similar, or some distinct locations blur together.
1 = revisits look unrelated, OR everything is one generic backdrop.
HARD RULE: if the system renders essentially one background for the whole story, score at
     most 2. Do not score merely-similar generic vegetation or a blank studio backdrop as
     a successful revisit.
Do not score above 3 just because two frames share a location NAME; the depicted place
     must actually match.

3. visual_quality
Craft and coherence within the declared style: anatomy, composition, rendering, freedom
     from artifacts and from garbled/nonsense lettering.
5 = clean and controlled.
3 = usable with visible defects.
1 = broken.

PROCEDURE
First identify the locations depicted across the sequence and compare frames corresponding
      to scene revisits before assigning scene_memory.
Then score each reported axis with a brief justification citing specific frames.

Return JSON only.
```

The structured output additionally contains a one-sentence justification for each axis, together with `revisit_analysis`, `degenerate`, and `note`. The main comparison uses Character Identity, Scene Memory, and Visual Fidelity from this evaluator.

Visual Memory and Aesthetic Quality evaluator

```
You are judging an AI visual story on TWO axes only.

You get: the declared STYLE, a short CAST/LOCATION spec, and the frames
ONE system produced (in order). Frame count may differ from the spec.
Do NOT assume frame k is shot k.

1. visual_memory
Did LATER frames reuse visual decisions that were first COMMITTED IN EARLIER
GENERATED FRAMES, not merely restated in the text?
Examples of memory: the exact mole placement, a specific shop-sign layout,
a unique knit pattern, the same window frame geometry after the camera left
and came back.
HARD RULES
- Matching a user reference photo is NOT memory. That is image conditioning.
  If the face is locked to a reference but locations/outfits/backgrounds
  are re-rolled or destroyed, score at most 2.
- Copying one panel into every frame is NOT memory. Set copy_collapse=1
  and score visual_memory = 1.
- Independent T2I that only follows the written description, with details
  re-rolling each frame, should score 2-3 even if each still looks pretty.
5 = later frames clearly remember earlier generated choices after a change
    of location or pose.
3 = mostly follows the written spec; little extra carry-over.
1 = new character/place each time, or one image repeated.

2. style_aesthetics
Beauty and craft INSIDE the declared style. If the spec asks for manga,
watercolour, greyscale, film grain, or hard flash, following that is GOOD.
Do NOT reward photorealism for its own sake.
MUST penalise: white speckle/glitter overlays, decorative polaroid/phone
chrome, garbled lettering, broken anatomy, one wallpaper reused as every
background.
5 = clean, attractive, on-style.
3 = usable, visible defects.
1 = broken or buried in artifacts.

Fill memory_analysis and style_analysis FIRST. Return JSON only.
```

Text Rendering evaluator

```
You are judging TEXT RENDERING.

STEP 1 -- list text frames
A frame counts if it has painted story text: shop signs, neon, price tags,
posters, magazines, menus, titles, speech balloons, handwritten notes.
Does NOT count: watermarks, site headers, Polaroid/Instagram chrome, EXIF
```

```
junk, pure texture that is not writing.
If ZERO frames qualify: set n_text_frames=0, text_frames="", text_seen=NONE,
and set text_rendering=0 and text_match=0. Do not invent a 1-5 score.

STEP 2 -- score only those frames (when n_text_frames >= 1)

1. text_rendering -- glyph quality on text frames only
Are the characters real and readable (correct Chinese / English strokes),
not melted fake radicals?
5 = almost all text frames are clean and readable.
3 = mixed: some words OK, others broken.
1 = writing is present but almost all garbled.

2. text_match -- wording on text frames only
If the spec names specific words, are they present and correct?
If the spec only wants generic signage, plausible on-theme text in the
right language is enough; random English on a Chinese street scores low.
5 = requested or clearly on-theme words are correct.
3 = right language / vibe, wording drifts.
1 = wrong language or nonsense.


Return JSON only.
```

F Baseline Implementation Details

F.1 Evaluation Paradigms

We evaluate two families of long-form generation systems. *Agent-Planned* methods decompose the story-level specification into a sequence of image-generation instructions and synthesize each image independently. This group includes StoryDiffusion, Qwen-Image, UniSpace, HiDream, Lance-3B, and Show-o2-7B. *Interleaved* methods instead receive the story-level instruction and generate text and images autoregressively within a shared multimodal trajectory. This group includes BAGEL-7B, SenseNova-U1.5, ILLUME-X, Emu3.5, SEED-Story, and UniLongGen*.

The two groups differ in how historical information is propagated. Agent-Planned systems rely on the information explicitly included in each image-generation instruction, whereas native interleaved systems can condition subsequent generation on the multimodal states already produced within the trajectory. All systems are evaluated using the same LongMIG-Bench story specifications and the evaluation protocols described in Appendix E.

F.2 Agent-Planned Baselines

For Agent-Planned baselines, each story is represented as a sequence of per-image generation instructions derived from the same underlying structured specification. We use GPT-6 as a shared planner for all Agent-Planned methods to decompose each story into an ordered shot plan and generate the corresponding per-image captions. This ensures that differences across methods primarily reflect their image-generation capabilities rather than differences in planning or prompt construction. Images are then generated independently at each step rather than being inserted into a shared autoregressive multimodal context. For the image-conditioned track, the provided user reference image is supplied whenever the corresponding method supports reference-conditioned generation.

We include strong general-purpose image generators as well as methods designed for multi-image or story generation. StoryDiffusion provides a representative story-generation baseline, while Qwen-Image, UniSpace, HiDream, Lance-3B, and Show-o2-7B provide strong per-image generation capability under the agent-planned formulation.

F.3 Native Interleaved Baselines

Native interleaved baselines are evaluated in their text-image generation mode, where the model receives the complete story-level instruction and produces a multimodal autoregressive trajectory. We evaluate BAGEL-7B, SenseNova-U1.5, ILLUME-X, Emu3.5, and SEED-Story. Unlike the Agent-Planned setting, generated images remain part of the evolving multimodal trajectory according to each model's native inference interface.

Table 6: **Efficiency across increasing generation horizons.** Ours and Full History use identical model weights and generation settings. Sec./Image is the mean wall-clock time per image over the first h images; Last Prefill is the AR prefill time of the h-th image.

	Sec./Image ↓		Last Prefill (s) ↓	
Horizon	Full History	Ours	Full History	Ours
8	27.8	26.4	6.8	4.9
16	49.3	37.6	20.7	9.8
24	73.8	46.5	41.4	14.0
32	102.9	54.9	69.0	19.4

For all methods, generated character-reference or auxiliary reasoning images that are not part of the final story are excluded from story-panel metrics. The evaluator receives only the resulting story panels together with the same ground-truth LongMIG-Bench specification.

F.4 UniLongGen Same-Backbone Reimplementation

To separate the effect of visual-memory management from backbone capability, we additionally reimplement the training-free context-curation strategy of UniLongGen on the same unified multimodal backbone used by our method, denoted as UniLongGen*. This comparison removes the large variation in image-generation quality and multimodal capability that otherwise exists across different unified models.

UniLongGen* follows the historical visual-context selection strategy of UniLongGen while retaining our underlying model weights. In contrast, our method constructs persistent visual references, predicts the current scene state, and retrieves a bounded set of scene-relevant historical generative states. The comparison therefore focuses on the effect of the visual-memory strategy under a shared backbone.

G Additional Efficiency and Scalability Analysis

We provide additional analysis of the computational behavior of our visual-memory mechanism as the generation horizon increases. Unless otherwise specified, Ours and Full History use identical model weights, decoding settings, and hardware; the only difference is whether historical generative visual states are selectively retrieved or all retained for image synthesis. Both settings preserve the complete semantic visual history for multimodal understanding.

G.1 Efficiency across Generation Horizons

We first evaluate eight long-form stories, each extended to 32 shots, at increasing generation horizons. Inference uses the same two-node setup for both policies, with one AR node (TP$=8$) and one image-generation node (TP$=8$), and 30 denoising steps for each generated image.

As shown in Table 6, the computational gap increases steadily with generation horizon. At 32 images, selective visual-memory retrieval reduces average generation time from 102.9 to 54.9 seconds per image, while the AR prefill time of the final image decreases from 69.0 to 19.4 seconds. The two policies are relatively close at short horizons, where fixed image-generation cost dominates, but diverge as historical visual states accumulate. Ours is faster on every story available under both policies at every horizon of 16 images or more.

G.2 Context and Latency Growth

To isolate the effect of historical visual conditioning, we additionally trace a fixed long-form generation under both policies. Figure 12 reports the image-synthesis context and per-image wall-clock time as the trajectory grows.

In this controlled trace, the image-synthesis context grows by approximately $1{,}546$ tokens per additional image with our routing policy, compared with $5{,}028$ tokens per image under Full History.

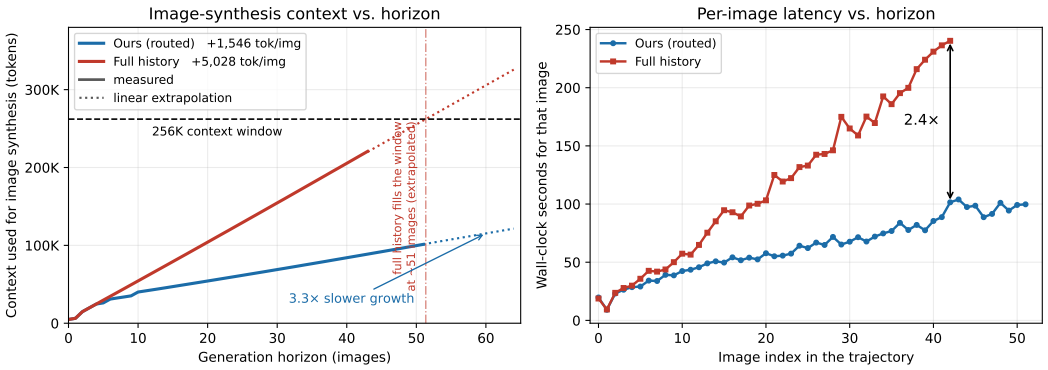


Figure 12: **Scaling of image-synthesis context and generation latency.** Left: selective visual-memory retrieval substantially slows context growth compared with Full History. Right: the corresponding per-image latency gap increases with generation horizon. Solid curves denote measured trajectories, while dotted extensions indicate linear extrapolation.

The corresponding wall-clock growth is approximately 1.5 versus 5.5 seconds per additional image. Thus, the marginal latency introduced by extending the trajectory grows substantially more slowly with selective visual-memory retrieval. Because the denoising schedule is fixed, the attention cost of image synthesis scales linearly with this context; at 32 images the context is 71.9K tokens under our policy versus 164.6K under Full History, a $2.3\times$ reduction.

H Failure Cases and Limitations

Despite the improvements in long-horizon identity and scene consistency, several failure modes remain. We summarize representative cases and clarify the scope of the conclusions supported by our experiments.

H.1 Fine-Grained Object Continuity

Our visual-memory mechanism operates primarily at the level of complete historical images and persistent visual references. It does not explicitly bind fine-grained object instances across the trajectory. Consequently, small carried objects, accessories, garment details, or objects placed within a scene can still disappear, change appearance, or be replaced after multiple generations.

H.2 Dependence on Explicit Scene-State Prediction

The current retrieval mechanism depends on an explicit discrete scene state. During training, scene annotations are obtained from the structured training data, whereas at inference time the model generates the scene state within its textual reasoning trajectory and the memory controller parses this prediction to determine the relevant historical visual states.

This creates a coupling between the model's generated reasoning format and the memory mechanism. If the model fails to emit a valid scene state, changes the output format, or predicts the wrong scene, retrieval may expose irrelevant historical visual states or fail to recover the intended establishing view. A representation-level scene state predicted directly from hidden features would remove this dependence on textual serialization and is a promising direction for future work.

H.3 Remaining Long-Horizon Generation Failures

Long-horizon consistency is improved but not perfect. We observe several recurring failure modes: identity drift in crowded multi-character scenes, imperfect recovery of fine scene details after very long gaps, repeated compositions in design-oriented sequences, and occasional degradation in ren-

dered text. Self-rollout reduces exposure mismatch but does not eliminate error accumulation when an early generation contains a substantial visual error.

These failures become more visible as the trajectory grows because generated outputs become part of the context for subsequent generation. Our method mitigates this accumulation by controlling which visual states are reused, but it cannot correct every error already encoded in the generated history.

Overall, our results demonstrate that visual-memory management substantially extends the usable horizon of native interleaved generation, but it does not eliminate all forms of autoregressive error accumulation. More structured object-level memory, representation-level scene tracking, and stronger adaptation to self-generated histories remain promising directions for further improving long-form multimodal generation.