Towards Unified Multimodal Interleaved Generation via Group Relative Policy Optimization

Ming Nie[1] **Chunwei Wang**[2] **Jianhua Han**[2] **Hang Xu**[2] **Li Zhang**[1*]

[1]School of Data Science, Fudan University [2]Noah's Ark Lab, Huawei

https://github.com/LogosRoboticsGroup/UnifiedGRPO

Abstract

Unified vision-language models have made significant progress in multimodal understanding and generation, yet they largely fall short in producing multimodal interleaved outputs, which is a crucial capability for tasks like visual storytelling and step-by-step visual reasoning. In this work, we propose a reinforcement learning-based post-training strategy to unlock this capability in existing unified models, without relying on large-scale multimodal interleaved datasets. We begin with a warm-up stage using a hybrid dataset comprising curated interleaved sequences and limited data for multimodal understanding and text-to-image generation, which exposes the model to interleaved generation patterns while preserving its pretrained capabilities. To further refine interleaved generation, we propose a unified policy optimization framework that extends Group Relative Policy Optimization (GRPO) to the multimodal setting. Our approach jointly models text and image generation within a single decoding trajectory and optimizes it with our novel hybrid rewards covering textual relevance, visual-text alignment, and structural fidelity. Additionally, we incorporate process-level rewards to provide step-wise guidance, enhancing training efficiency in complex multimodal tasks. Experiments on MMIE and InterleavedBench demonstrate that our approach significantly enhances the quality and coherence of multimodal interleaved generation.

1 Introduction

In recent years, rapid progress in visual understanding and generation has driven a growing trend towards unifying these capabilities within a single multimodal framework. Unified vision-language models aim to perform both understanding tasks (*e.g.*, visual question answering) and generation tasks (*e.g.*, text-to-image generation) within one model, representing a pivotal direction in the evolution of vision-language models. This paradigm has seen significant advancement, enabled by the availability of large-scale image-text paired datasets and the scaling of multimodal model architectures. Recent models such as Show-O, VILA-U, and ILLUME [34, 32, 28] demonstrate strong generalization across diverse benchmarks, underscoring the potential of unified frameworks as general-purpose AI systems for integrated vision-language reasoning and content creation.

Despite their impressive performance, most unified models still struggle to generate multimodal interleaved contents, which is a critical capability for enabling fine-grained reasoning, step-by-step explanation, and context-aware multimodal synthesis. During inference, these models typically produce either text-only or image-only outputs, constrained by explicit or implicit modality control mechanisms. This limitation arises primarily from the lack of fine-grained supervision and training data to guide dynamic modality transitions. Consequently, current models often default to generating single-modality outputs, failing on tasks that require tightly coupled multimodal sequences, such

*Li Zhang (lizhangfd@fudan.edu.cn) is the corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

as visual dialogue or sequential storytelling. Addressing this gap is essential for realizing the full potential of unified models in seamless multimodal reasoning and generation.

To this end, we propose a post-training strategy that unlocks multimodal interleaved generation without requiring large-scale high-quality interleaved data. We hypothesize that unified models inherently possess fundamental multimodal generation capabilities acquired during pre-training and supervised fine-tuning (SFT), and that minimal interleaved supervision is sufficient to activate them. Based on this hypothesis, we first construct a hybrid dataset comprising a small amount of interleaved text-image sequences to expose the model to multimodal generation patterns, while incorporating limited SFT data to retain its pretrained strengths in multimodal understanding and conventional text-to-image generation. After this warm-up stage, the unified model is capable of generating basic multimodal interleaved contents conditioned on instructions. However, the outputs often suffer from weak cross-modal alignment, reflecting limited consistency and coherence between text and image.

To further improve multimodal interleaved generation, we propose a reinforcement learning algorithm that formulates generation as a sequential decision-making process and extends Group Relative Policy Optimization [21] (GRPO) to the multimodal setting. While GRPO has been widely explored for text-only modalities, it struggles with multimodal outputs due to challenges like modality switching and hybrid reward attribution. To address this, we propose a unified policy optimization framework that jointly models text and image generation under a single decoding trajectory. This approach enables modality-aware decisions and leverages a shared training objective, preserving GRPO's advantages while adapting to multimodal structures. Specifically, we define a hybrid group-wise reward signal consisting of three key components: (i) a textual reward that evaluates the quality and relevance of the generated text conditioned on the input prompt; (ii) a visual and multimodal reward that jointly assesses image quality and image-text consistency; and (iii) a format reward that promotes structural fidelity by penalizing violations in the expected interleaved output format. In addition, we incorporate process-level reward modeling by assigning intermediate rewards at the interleaved generation steps. This design provides more granular and timely feedback throughout the generation process, which significantly improves the efficiency and effectiveness of policy learning. Such fine-grained supervision is particularly advantageous in complex multimodal generation tasks, where step-wise feedback helps the model perform autoregressive interleaved generation more effectively than relying solely on outcome rewards. We evaluate our approach on two dedicated multimodal interleaved generation benchmarks, namely MMIE [33] and InterleavedBench [14], and demonstrate its effectiveness in enabling unified models to generate coherent, high-quality interleaved outputs.

To summarize, our main contributions are as follows: **(i)** We introduce a warm-up stage that unlocks the model's latent capability for interleaved text-image generation using only a small amount of curated interleaved data; **(ii)** We propose a unified policy optimization framework that supports autoregressive generation across both text and image modalities, enabling seamless modality switching within a single decoding trajectory; **(iii)** We design a group of hybrid rewards that supervises multimodal generation from multiple aspects, and further enhance learning via process-level rewards that provide step-wise guidance; **(iv)** We conduct extensive experiments on two challenging benchmarks, MMIE and InterleavedBench, demonstrating the effectiveness and efficiency of our approach compared to existing unified models.

2 Related work

2.1 Unified understanding and generation

Recently, an increasing number of studies [34, 40, 26, 27, 32, 4, 31] have focused on unified models for both multimodal understanding and generation. Show-o [34], TransFusion [40] and Janus-Pro [4] adopt hybrid diffusion-autoregressive strategies, while Emu2 [26] employs a fully autoregressive architecture that predicts the next multimodal element via classification for text and regression for visual embeddings. Chameleon [27], VILA-U [32], and other works [28, 31] tokenize images and interleave them with text tokens, enabling joint reasoning and autoregressive generation across modalities. Despite these efforts, current unified models often fail to support fine-grained interleaved generation of text and images, largely due to the scarcity of high-quality multimodal interleaved data. This limitation restricts their applicability in tasks that require coherent and context-aware multimodal dialogue and reasoning.

2.2 Multimodal interleaved generation

Multimodal learning has rapidly advanced, especially in the integration of text and image modalities. Early models such as DALL·E [19] and Stable Diffusion [16] demonstrated impressive capabilities in generating images from text prompts, but they were primarily designed for unidirectional generation, either text-to-image or image-to-text, without supporting interleaved multimodal outputs. Recently, large vision-language models (LVLMs) have begun to tackle this limitation by enabling interleaved generation, where text and images are jointly generated within a single autoregressive sequence. Approaches such as Chameleon [27], GILL [9], Anole [5] and ARMOR [24], have pushed the boundaries of token-level multimodal modeling. These models adopt unified generation frameworks that allow seamless alternation between modalities, supporting more natural and interactive multimodal communication. However, a key challenge lies in the scarcity of large-scale, fine-grained multimodal supervision data, which limits the full potential of interleaved generation. To address this, we propose a post-training strategy that activates the model's multimodal generation capability without relying on extensive high-quality training data.

2.3 Policy optimization in multimodal models

Policy optimization [15, 18, 21, 8, 22, 30] has become a pivotal component in aligning large language and vision-language models with human intent. Reinforcement learning from human feedback (RLHF) [15], exemplified by methods such as PPO [20] and DPO [18], has proven effective in improving text-only generation by leveraging reward signals derived from preference data or learned reward models. Recently, Group Relative Policy Optimization (GRPO) [21] has demonstrated promising results in the post-training of large language models, particularly in domains requiring complex reasoning such as mathematics [8, 22, 30]. Its group-wise comparative reward mechanism offers improved sample efficiency and stability, making it a compelling foundation for extension into multimodal interleaved generation. However, extending policy optimization to multimodal settings remains a substantial challenge. First, existing methods [41, 29, 3] are often limited to text-only modalities or apply separate optimization procedures for text and image components, which hinders the development of unified models capable of seamless interleaved generation. Second, most approaches rely on outcome-oriented, end-of-sequence rewards, which are inherently sparse and fail to capture the fine-grained, step-wise alignment required for complex multimodal tasks. In this work, we propose a unified policy optimization algorithm, which treats multimodal generation as a single coherent decision-making process for enhanced multimodal policy optimization.

3 Methodology

In this section, we first present the preliminary formulation of unified multimodal models in 3.1. We then introduce a warm-up training scheme to equip the model with the capability of generating interleaved text-image outputs while preserving its pretrained knowledge in Sec. 3.2. Next, in Sec. 3.3, we describe our proposed GRPO training framework based on hybrid reward signals tailored for multimodal generation. Finally, we outline the training details in Sec. 3.4.

3.1 Preliminary

Unified multimodal models seek to integrate vision and language by leveraging a shared architecture capable of jointly modeling textual and visual modalities in both understanding and generation tasks. Given a multimodal input sequence $X = \{x_1, x_2, \ldots, x_n\}$, where each token x_i can correspond to either a textual or visual token, the model learns to autoregressively generate an output sequence $Y = \{y_1, \ldots, y_m\}$, which is also capable of comprising both modalities. The generation process in unified models is typically modeled as an autoregressive probability distribution over the output tokens:

$$p(Y|X) = \prod_{t=1}^{m} p(y_t|X, y_{<t}; \theta), \tag{1}$$

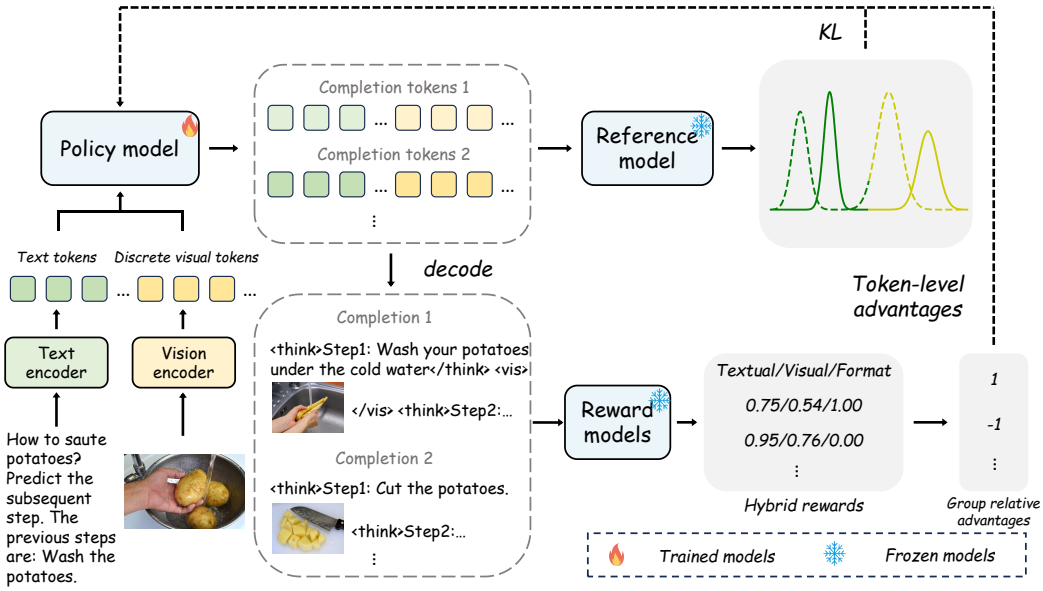

Figure 1: Overview of our reinforcement fine-tuning framework. Multimodal tokens are autoregressively generated and decoded into completions, with token probabilities used to compute KL divergence along a single trajectory. Hybrid rewards are assigned to each completion, and token-level group relative advantages are calculated to guide policy optimization along with KL regularization.

where θ denotes the model parameters, and the conditional generation at each step relies on both the input sequence X and the previously generated tokens $y_{<t}$. The unified model is generally optimized using a standard language modeling objective covering multimodal sequences:

$$\mathcal{L} = -(\sum_{t=1}^{T} \mathbb{I}_{txt}(t) log P(y_t|x, y_{<t}) + \sum_{t=1}^{T} \mathbb{I}_{img}(t) log P(y_t|x, y_{<t})). \tag{2}$$

This framework allows for flexible training across various tasks simultaneously, such as captioning, image generation, and more complex multimodal reasoning. However, due to the lack of high-quality training data that explicitly supervises interleaved text-image generation, current unified models often struggle to realize their potential for multimodal generation, leading to limited and inconsistent performance in such tasks.

3.2 Unlocking interleaved generation

To equip unified models with interleaved generation and multimodal reasoning capabilities, we propose a warm-up training scheme that activates new functionalities while preserving existing strengths. Unified models are pretrained on large-scale multimodal data comprising both text and images, theoretically equipping them for multimodal interleaved generation. However, they often struggle with smoothly transitioning between modalities and aligning context across text and image. Direct fine-tuning on novel tasks may substantially impair the model's pretrained competencies and induce catastrophic forgetting. To address this, we introduce a hybrid data-based warm-up phase that bridges pretraining and task-specific fine-tuning. In this stage, we incorporate a large corpus of high-quality supervised fine-tuning (SFT) data, including either curated human-labeled samples or distilled outputs from strong multimodal LLMs. We further introduce interleaved text-image data to expose the model to multimodal generation patterns, while blending in purely textual data to preserve and reinforce its language understanding and generation capabilities. This strategy helps unlock interleaved generation abilities without compromising the model's original strengths.

3.3 Reinforcement fine-tuning

Once the model is warmed up, we proceed to a GRPO-based optimization phase to further enhance generation quality and cross-modal alignment. During reinforcement fine-tuning, we apply the GRPO algorithm guided by a hybrid supervised signal comprising both rule-based metrics and model-based reward, as shown in Figure 1.

Unified policy for multimodal generation. To enhance interleaved text-image generation and multimodal reasoning, we adapt Group Relative Policy Optimization (GRPO) [21], an efficient reinforcement learning algorithm originally proposed for text-only LLMs. GRPO estimates token-level advantages by conducting intra-group comparisons among generated responses conditioned on the same query. Given an input instance X, a behavior agent $\phi_{\theta_{old}}$ samples a batch of G candidate responses $\{Y_i\}_{i=1}^{G}$. However, previous applications of GRPO have been limited to optimizing text-only policies. To extend it to unified models capable of generating both text and images, we treat the multimodal generation as a single decision process: $Y_i = \{y_1^{txt}, \ldots, y_k^{txt}, y_{k+1}^{img}, \ldots, y_m^{img}\}$, and adapt the GRPO framework accordingly. The GRPO optimization objective incorporates a clipped surrogate loss term augmented with a KL-penalty to ensure stable policy updates, formulated as:

$$\mathcal{L}_{GRPO}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|Y_i|}\sum_{t=1}^{|Y_i|} min[\mathbb{D}_{uni}(t)\hat{A}_{i,t}, clip(\mathbb{D}_{uni}(t), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}], \tag{3}$$

where $\mathbb{D}_{uni}$ represents the unified KL divergence computed per token across modalities:

$$\mathbb{D}_{uni}(t) = \frac{\pi_\theta(y_{i,t}^{txt}|X_i, y_{i,\leq t})}{\pi_{\theta_{old}}(y_{i,t}^{txt}|X_i, y_{i,\leq t})}\mathbb{I}_{t\leq k}(t) + \frac{\pi_\theta(y_{i,t}^{img}|X_i, y_{i,\leq t})}{\pi_{\theta_{old}}(y_{i,t}^{img}|X_i, y_{i,\leq t})}\mathbb{I}_{t>k}(t). \tag{4}$$

The advantage $\hat{A}_{i,t}$ for the i-th response at time step t is determined by normalizing the rewards obtained across the response group, according to:

$$\hat{A}_{i,t} = \frac{r(X, Y_i) - mean(\{r(X, Y_1), ..., r(X, Y_G)\})}{std(\{r(X, Y_1), ..., r(X, Y_G)\})}. \tag{5}$$

The hyperparameter ϵ controls the permissible deviation from the reference policy. The clipping function constrains the policy ratio within a specified interval, preventing overly large updates. This regularization mechanism enhances training stability and reduces the risk of performance collapse due to aggressive policy shifts.

Hybrid reward signal. Formally, a group of reward models assign a scalar score to each response, producing a set of G rewards $r = \{r_1, \ldots, r_G\}$ corresponding to the G candidates. To guide the optimization of the unified model, we design a hybrid reward signal that integrates multiple task-specific objectives. The hybrid reward aims to jointly enhance the quality of both generated text and images, as well as ensure output format consistency.

To better model and optimize the interleaved multimodal outputs, we begin with a probabilistic decomposition of the joint distribution over input prompt X, generated text Y_{txt}, and generated image Y_{img}. We decompose the joint probability as:

$$p(Y_{txt}, Y_{img}|X) = p(Y_{txt}|X)p(Y_{img}|X, Y_{txt}). \tag{6}$$

From this perspective, we design our hybrid reward to align with the two conditional components: The textual reward r_t evaluates $p(Y_{txt}|X)$, assessing the relevance and coherence of the generated text given the prompts. The visual rewards r_v targets $p(Y_{img}|X, Y_{txt})$, measuring the image's quality and its alignment with the text and prompt context. Furthermore, we incorporate a format reward r_f to regularize the structure of the generated content. Specifically, we leverage special tokens `<think>` and `<vis>` to explicitly separate different modalities in the content sequence, guiding the model to adhere to a consistent and interpretable format. This formatting strategy is illustrated in Figure 1. The textual relevance reward r_t is assessed by [35] and the visual reward is assessed by ImageReward [36]. To this end, our hybrid reward function is defined as follows:

Method	Params	MMIE			
		Situational analysis	Project-based learning	Multi-step reasoning	AVG
MiniGPT-5 [39]	7B	47.63	55.12	42.17	50.92
EMU-2 [26]	37B	39.65	46.12	50.75	45.33
GILL [9]	7B	46.72	57.57	39.33	51.58
Anole [5]	7B	48.95	59.05	51.72	55.22
Ours (QLoRA)	7B	53.12	60.34	53.28	57.27
Ours	7B	**56.87**	**62.28**	**54.31**	**59.50**

Table 1: Comparison with existing methods on MMIE. The proposed method significantly improves the generation quality of unified models, producing interleaved text–image outputs that are both coherent and aligned with instructions.

$$r(X, Y_i) = r_t(X, Y_i) + r_v(X, Y_i) + r_f(X, Y_i). \tag{7}$$

Process-level reward. While outcome supervision provides a single reward at the end of each output, this sparse feedback may be insufficient and inefficient for complex multimodal generation tasks. To address this, we additionally incorporate process-level supervision, which offers intermediate rewards at the end of each modality step. Formally, given a set of sampled outputs $\{Y_i\}_{i=1}^{G}$, a process reward model assigns a sequence of rewards to each output: $R = \{\{r_1^{index(1)}, ..., r_1^{index(K_1)}\}, ..., \{r_G^{index(1)}, ..., r_G^{index(K_G)}\}\}$, where $index(j)$ denotes the end token index of the j-th step, and K_i is the total number of modality steps in the i-th output. Each step reward is then normalized using the overall mean and standard deviation across all samples. To compute token-level advantages, we accumulate the normalized rewards from all subsequent steps: $\hat{A}_{i,t} = \sum_{index(j) \geq t} \hat{r}_i^{index(j)}$. The final policy is then optimized by maximizing the GRPO objective using these token-level advantages.

3.4 Training details

We implement VILA-U [32] as our foundation unified model. Thanks to its unified pretraining paradigm, which jointly learns multimodal comprehension and text-to-image generation, the model inherently possesses the potential for multimodal output. Our approach builds on this capability and unlocks interleaved generation with only minimal additional data. During the warm-up stage, we collect 0.3M text-image paired samples with interleaved outputs from ActivityNet [1], GenHowTo [23] and OpenStory++ [37]. To preserve the model's original multimodal understanding and generation abilities, we further incorporate 1M multimodal understanding samples from EMOVA [2] and 1M text-to-image generation samples from JourneyDB [25]. For the GRPO stage, we curate a dataset of 0.1M samples (from the same sources as in the warm-up stage) focusing on visual storytelling and multimodal interleaved reasoning to facilitate effective policy optimization.

4 Experiment

4.1 Experimental setup

Data preparation. During the warm-up stage, we collect 0.3M interleaved text-image samples from ActivityNet [1], GenHowTo [23], and OpenStory++ [37]. The data covers diverse multimodal scenarios, including temporally grounded actions, step-by-step state transitions, and narrative storytelling, providing rich interleaved patterns for training. **(i)** ActivityNet [1] is a large-scale benchmark for human activity understanding, covering 203 activity classes with about 137 untrimmed videos per class and 1.41 activity instances per video (849 hours total). We use its temporal dense captions [10] to describe fine-grained action segments. For each localized clip, a duration-adaptive number of

Method	Params	InterleavedBench					
		Text quality	Perceptual quality	Image coherence	TIC	Helpfulness	AVG
MiniGPT-5 [39]	7B	1.22	2.45	1.62	2.03	1.77	1.82
EMU-2 [26]	37B	1.26	2.28	1.89	1.34	1.64	1.68
GILL [9]	7B	0.75	3.21	2.25	1.53	1.48	1.84
Ours (QLoRA)	7B	2.33	3.39	3.01	2.75	2.85	2.87
Ours	7B	**2.86**	**3.58**	**3.25**	**3.02**	**2.94**	**3.13**

Table 2: Comparison with existing methods on InterleavedBench, where our method demonstrates superior performance across multiple evaluation metrics.

keyframes is extracted to capture visual details, and these keyframes with their captions form interleaved training samples. **(ii)** GenHowTo is a large-scale dataset automatically constructed with 200K image triplets and corresponding textual descriptions. Each triplet contains temporally ordered frames depicting an object's initial state, the modifying action, and the resulting new state, effectively modeling cause-and-effect visual transitions. These structured triplets naturally support interleaved text-image training for step-by-step visual reasoning. We leverage both the provided action and state prompts, along with their aligned images, to construct coherent interleaved text-image samples for visual reasoning tasks. **(iii)** OpenStory++ [37] is a comprehensive dataset emphasizing narrative continuity with instance-level visual segmentation. It extracts keyframes from videos, filters them by aesthetic quality, and generates refined captions via an LLM-based annotator to ensure coherence. The resulting keyframe–caption pairs form interleaved text-image sequences for visual storytelling. More detailed data construction process and illustration are provided in the appendix.

Implementation details. As described previously, we adopt VILA-U [32] as our foundation unified model. The language model, vision encoder, and visual decoder are all implemented following VILA-U's original architecture and configuration. All images are resized to a fixed resolution of 256 $\times$ 256 before being fed into the model. For image generation, we employ classifier-free guidance with a guidance scale of 3 to improve output fidelity. In GRPO stage, the number of generation G is set to 4 and we train the model for 3k steps. All full-scale experiments are conducted using 32 NVIDIA A100 GPUs. To further demonstrate the compatibility and efficiency of our method, we also implement a lightweight version using QLoRA, achieving competitive performance with significantly lower memory and compute requirements.

Evaluation. We evaluate the model's capability for interleaved multimodal generation on two dedicated benchmarks: MMIE[33] and InterleavedBench[14]. MMIE consists of 20K carefully curated multimodal queries, covering 3 main categories, 12 domains, and 102 subfields, including mathematics, coding, physics, literature, health, and the arts. It introduces a robust automated evaluation protocol using a scoring model fine-tuned on human-annotated data with systematic criteria, aiming to reduce bias and enhance evaluation reliability. InterleavedBench comprises 815 instances across 10 real-world use cases. It defines five key evaluation dimensions: text quality, perceptual quality, image coherence, text-image coherence, and overall helpfulness, offering a comprehensive and fine-grained assessment of interleaved generation performance. By encompassing detailed evaluation criteria, InterleavedBench provides a robust framework for measuring a model's ability to generate coherent, visually aligned, and contextually meaningful interleaved outputs.

4.2 Main results

We compare our approach against existing unified models on these challenging benchmarks to evaluate its effectiveness in interleaved multimodal generation. As shown in Table 1, our approach significantly improves the ability of unified models to produce coherent and instruction-aligned text-image outputs. These improvements demonstrate that our method not only strengthens the generation quality under limited supervision, but also pushes the performance upper bound of current unified models on interleaved multimodal generation benchmarks (59.5% on MMIE). In particular, our method demonstrates a clear advantage on the situational analysis task, which evaluates a model's ability to perform visual storytelling based on given textual and visual prompts. Our model achieves

Warm-up	GRPO	MMIE	InterleavedBench
✗	✗	-	0.51
✓	✗	53.31	1.97
✓	✓	59.50	3.13

Table 3: Effects of warm-up stage and GRPO. The warm-up stage enables to generate interleaved multimodal outputs. Building on the warm-up initialization, the proposed GRPO-based policy optimization substantially improves the quality of interleaved generations.

Hybrid rewards			Process reward	MMIE	InterleavedBench
r_f	r_t	r_v			
✓	✗	✗	✗	53.56	2.05
✓	✓	✗	✗	54.62	2.30
✓	✓	✓	✗	57.83	2.79
✓	✓	✓	✓	59.50	3.13

Table 4: Hybrid reward components ablation. Beginning with the format reward r_f, we gradually integrate the textual, visual and process-level rewards. The results demonstrate that the inclusion of complementary reward signals progressively enhances interleaved generation quality.

a score of 56.87%, outperforming Anole by over 10%. This advanced performance highlights the effectiveness of our interleaved generation strategy in handling complex, context-rich multimodal reasoning scenarios. We also report our results on InterleavedBench, as shown in Table 2. Our method outperforms existing unified models, achieving superior performance with a gain of 1.29 over GILL, demonstrating the robustness and generalization ability of our approach.

4.3 Ablation study

Effects of warm-up stage and GRPO. We first investigate the impact of GRPO as well as warm-up stage in Table 3. Empirical results demonstrate that the warm-up stage effectively unlocks the unified model's capability to generate interleaved multimodal outputs (53.31% on MMIE and 1.97 on InterleavedBench). Prior to this stage, existing unified models were unable to produce such interleaved generations, and thus failed to yield valid results on benchmarks like MMIE. Notably, InterleavedBench includes a subset of tasks that do not require multimodal outputs, which allows baseline models to achieve a marginal score (0.51). Building upon the warm-up initialization, our proposed GRPO-based policy optimization significantly enhances the quality of interleaved outputs, leading to consistent improvements across both benchmarks (+6.19% and +1.16). This two-stage strategy proves essential for enabling instruction-following, coherent, and structurally aligned text-image sequences.

Rewards components ablation. We then take a further step to investigating the components of rewards in the GRPO training paradigm. Starting from the format reward r_f, we progressively incorporate the textual reward r_t, visual reward r_v, and finally the process-level reward. As shown in Table 4, the format reward alone does not lead to significant performance gains compared to the warm-up stage. However, it plays an essential role in providing explicit supervision for modality switching, which is crucial for stable interleaved generation. Adding the textual reward yields a steady improvement of 1.06% on MMIE and 0.25 on InterleavedBench, while the visual reward further boosts performance by 3.01% and 0.49, respectively. Finally, integrating the process-level reward results in the best performance, raising the scores to 59.50% on MMIE and 3.13 on InterleavedBench.

Method	ILO-Uni.	Params	Understanding				Generation		
			MME-P [13]	MMvet [38]	SEEDBench-img [11]	POPE [12]	#Train images	Image res.	Score
VILA-U [32]	✗	7B	1401.8	33.5	59.0	85.8	15M	384	0.42
SEED-X [6]	✗	17B	1435.7	-	-	84.2	158M	-	0.49
TokenFlow-XL [17]	✗	13B	1545.9	40.7	68.7	86.8	60M	384	0.55
Janus-Pro [4]	✗	7B	1516.7	45.1	70.1	78.9	72M	384	0.80
Chameleon [27]	✓	7B	153.1	8.3	30.5	19.4	1.4B	512	0.39
ARMOR [24]	✓	8B	1619.4	53.1	74.8	87.7	5M	256	0.37
Ours	✓	7B	1425.2	32.8	59.2	85.1	16M	256	0.46

Table 5: Comparison with existing methods on visual understanding and generation benchmarks. ILO-Uni. denotes "Unified models with interleaved text-image output". Our approach maintains comparable performance to baseline models.

KL	G	r_v	MMIE	InterleavedBench
✗	2	ImageReward [36]	53.04	1.72
✓	2	ImageReward	55.14	2.27
✓	4	ImageReward	59.50	3.13
✓	4	CLIP-score	56.94	2.58

Table 6: Aspects of GRPO hyper-parameter implementation. We first analyze the effect of the KL-penalty term by duplicating its application during training. Results show that maintaining a KL constraint stabilizes optimization and mitigates policy drift from the initialization, underscoring its critical role. We then ablate the number of generations G to evaluate its impact on learning dynamics, and finally, we compare different formulations of the visual reward function.

These results demonstrate the effectiveness of each component and highlight the importance of fine-grained, multi-dimensional reward signals in training high-quality interleaved generation models.

Unified understanding and generation. Additionally, we evaluate our model's performance on standard multimodal understanding and generation tasks (GenEval [7]) to ensure that our method does not lead to performance degradation or catastrophic forgetting of existing unified capabilities. As shown in Table 5, our approach achieves comparable results to the baseline model across these tasks, indicating that both the warm-up stage and our GRPO-based training strategy are effective in preserving the model's original strengths. Furthermore, we compare our method with existing unified models that are capable of generating multimodal interleaved outputs. Results show that our approach achieves competitive performance, highlighting its advantage in enabling high-quality and instruction-aligned interleaved generation without sacrificing general multimodal capabilities.

Aspects of GRPO hyper-parameter implementation. We conduct a series of ablation studies to examine the impact of key hyper-parameter choices in our GRPO framework, as shown in Table 6. We begin by analyzing the role of the KL-penalty term, where we duplicate its application during training. Results show that enforcing KL divergence stabilizes training and prevents the model from drifting too far from the initial policy, confirming its importance.

We further ablate the number of generations G for each prompt during training. Increasing the number of sampled generations from 2 to 4 results in a notable performance improvement (+4.36% and +0.86), indicating that generating a more diverse set of candidates facilitates more accurate group-wise reward estimation and provides richer learning signals. However, due to computational limitations, we are unable to explore larger generation numbers. It is worth noting that GRPO is particularly resource-intensive in the multimodal setting, as it requires generating both textual and visual outputs during each rollout. This underscores the importance of developing more efficient and scalable post-training strategies for multimodal models in future research.

Figure 2: Visualizations of multimodal interleaved generation. Qualitative examples illustrate the model's capacity to produce coherent interleaved outputs, smoothly transitioning between text and image modalities within a unified generation process.

Lastly, we compare different choices of visual reward functions r_v. Specifically, we find that using ImageReward provides more accurate and consistent supervision signals for image quality than CLIP-score, reinforcing the value of tailored reward models in multimodal training.

Visualizations. Figure 2 showcases qualitative examples of interleaved generation, illustrating the model's ability to seamlessly alternate between text and image modalities. Notably, the visual outputs align well with the surrounding textual content, reflecting the effectiveness of our policy optimization strategy in producing fine-grained, semantically consistent interleaved content. More visualizations and failure case studies are included in the appendix.

5 Conclusion and discussions

Conclusion. We propose an effective training strategy to enable unified vision-language models to perform high-quality interleaved multimodal generation. Through a warm-up stage with limited data, we unlock the model's latent multimodal generation capabilities. We further introduce a unified policy optimization framework based on GRPO with hybrid and process-level rewards, which formulates the multimodal generation process as a single trajectory from a reinforcement learning perspective. Experiments on MMIE and InterleavedBench demonstrate superior performance over existing unified models while preserving general multimodal capabilities, paving the way for more versatile and controllable multimodal generation systems.

Discussion on limitations. Despite the significant performance gains on multimodal interleaved benchmarks, we observe that our method does not bring notable improvements on generalized multimodal understanding and text-to-image generation tasks. We attribute this to the underlying base model, which largely determines performance limits, with GRPO mainly serving to better align text and image content rather than enhancing general capabilities. Future work may benefit from stronger unified architectures and broader reward designs to further enhance model capability.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No. 62376060).

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 6, 21

[2] Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. Emova: Empowering language models to see, hear and speak with vivid emotions. *arXiv preprint*, 2024. 6

[3] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint*, 2025. 3

[4] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint*, 2025. 2, 9

[5] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint*, 2024. 3, 6

[6] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint*, 2024. 9

[7] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NIPS*, 2023. 9

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*, 2025. 3

[9] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. In *NIPS*, 2023. 3, 6, 7

[10] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 6, 21

[11] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024. 9

[12] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint*, 2023. 9

[13] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, 2024. 9

[14] Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. In *EMNLP*, 2024. 2, 7

[15] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. Training language models to follow instructions with human feedback. In *NIPS*, 2022. 3

[16] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint*, 2023. 3

[17] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint*, 2024. 9

[18] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NIPS*, 2023. 3

[19] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 3

[20] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint*, 2017. 3

[21] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint*, 2024. 2, 3, 5

[22] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint*, 2025. 3

[23] Tomáš Souček, Dima Damen, Michael Wray, Ivan Laptev, and Josef Sivic. Genhowto: Learning to generate actions and state transformations from instructional videos. In *CVPR*, 2024. 6, 21

[24] Jianwen Sun, Yukang Feng, Chuanhao Li, Fanrui Zhang, Zizhen Li, Jiaxin Ai, Sizhuo Zhou, Yu Dai, Shenglin Zhang, and Kaipeng Zhang. Armor v0. 1: Empowering autoregressive multimodal understanding model with interleaved multimodal generation via asymmetric synergy. *arXiv preprint*, 2025. 3, 9

[25] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: a benchmark for generative image understanding. In *NIPS*, 2023. 6

[26] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024. 2, 6, 7

[27] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint*, 2024. 2, 3, 9

[28] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint*, 2024. 1, 2

[29] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint*, 2025. 3

[30] Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, Xuchen Song, et al. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint*, 2025. 3

[31] Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable multi-modal generators. *arXiv preprint*, 2024. 2

[32] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint*, 2024. 1, 2, 6, 7, 9

[33] Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. In *ICLR*, 2025. 2, 7

[34] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint*, 2024. 1, 2

[35] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. In *CVPR*, 2025. 5

[36] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023. 5, 9

[37] Zilyu Ye, Jinxiu Liu, Ruotian Peng, Jinjin Cao, Zhiyang Chen, Yiyang Zhang, Ziwei Xuan, Mingyuan Zhou, Xiaoqian Shen, Mohamed Elhoseiny, et al. Openstory++: A large-scale dataset and benchmark for instance-aware open-domain visual storytelling. *arXiv preprint*, 2024. 6, 7, 21

[38] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2024. 9

[39] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint*, 2023. 6, 7

[40] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint*, 2024. 2

[41] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint*, 2025. 3

NeurIPS Paper Checklist

BEGIN INSTRUCTIONS

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We make the main claims clearly in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation of this paper in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We discuss the experimental setup in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code would be released after submission and review.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details of training and testing are specified in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Experiments in this paper are conducted in multiple iterations, yielding stable results. We reported the average of these results for consistency and reliability.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computational cost and resources are specified in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Broader Impacts are discussed in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We follow the safeguards of base unified model VILA-U.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models and datasets used in this paper have been properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM is an original component of the core methods in this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Data preparation

During the warm-up stage, we collect 0.3M text-image paired samples with interleaved outputs from ActivityNet [1], GenHowTo [23] and OpenStory++ [37]. The data comprises diverse multimodal scenarios, including temporally grounded action descriptions, step-by-step action-state pairs, and narrative visual storytelling. These samples are organized into interleaved text-image sequences to expose the model to rich multimodal generation patterns during training.

ActivityNet. ActivityNet [1] is a large-scale video benchmark designed for human activity understanding. It covers a broad spectrum of complex daily activities, featuring 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video, totaling approximately 849 hours of video content. We utilize the temporal dense captions that describe fine-grained action segments in each video [10]. For each temporally localized clip, we extract a set of representative keyframes to capture fine-grained visual details. The number of keyframes is adaptively determined based on the duration of the action clip. These keyframes, along with their corresponding dense captions, are subsequently used to construct interleaved training samples.

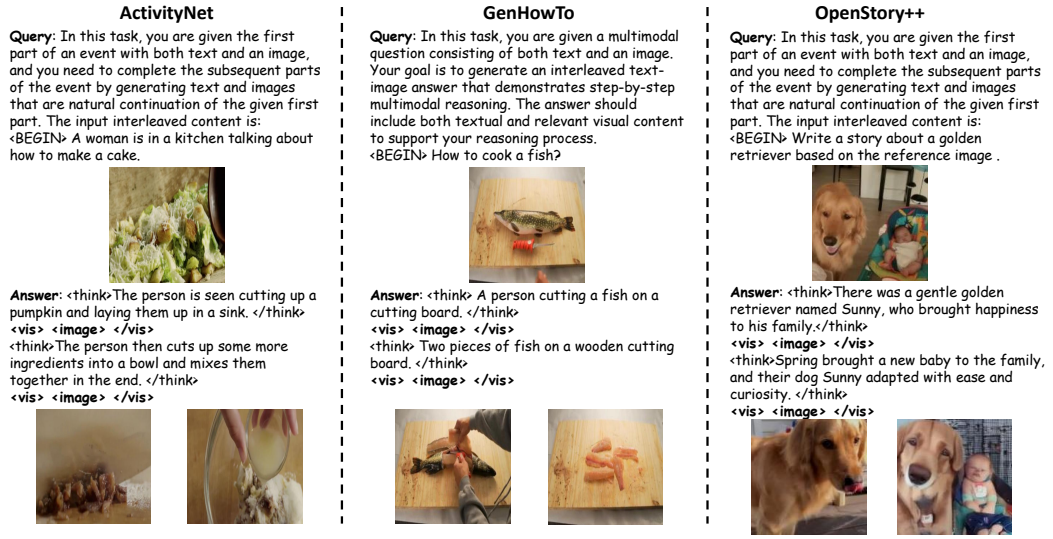

Figure 3: Illustration of data preparation process during the warm-up stage.

GenHowTo. GenHowTo is a large-scale dataset automatically constructed with 200K image triplets and corresponding textual descriptions. Each triplet consists of temporally ordered video frames depicting (i) the initial state of an object, (ii) the action that modifies the state, and (iii) the resulting new state of the object. These structured triplets, along with their descriptions, naturally align with interleaved text-image training for modeling step-by-step visual reasoning. We utilize the provided action prompts and state prompts, along with their corresponding images, to construct a series of interleaved text-image samples for visual reasoning tasks, as shown in Figure 3.

OpenStory++. OpenStory++ is a comprehensive dataset designed to emphasize narrative continuity around key instances, featuring instance-level visual segmentation. It processes video content by extracting keyframes, evaluating them for aesthetic quality, and generating descriptive captions using BLIP2. These captions are then refined by a Large Language Model (LLM) to ensure coherence and maintain narrative flow. The resulting keyframe-caption pairs are used to build interleaved text-image sequences that support visual storytelling, as shown in Figure 3.

B Failure cases analysis

While our method demonstrates strong performance on interleaved multimodal generation tasks, we observe several failure cases that highlight current limitations, as demonstrated in Figure 4. First, in some complex reasoning scenarios, the model may generate hallucinated visual content that does not

Figure 4: Failure cases analysis about hallucinated visual content.

faithfully correspond to the textual context, indicating limitations in cross-modal alignment. Second, the visual elements produced by the model occasionally fail to match the style or appearance of the input image, resulting in stylistic inconsistency across modalities. These issues suggest future work could benefit from improved reward modeling that better captures factual consistency, reasoning traceability, and structural fidelity in interleaved generation.

C Broader impacts

The proposed work has the potential to greatly enhance the capability of unified models for multimodal interleaved generation. Therefore it could be beneficial for applications such as multimodal reasoning and visual storytelling. It could also be used to provide a better experience of multimodal interleaved interaction between human and AI system. As for the potential negative impacts, there is a risk that the technology could be misused to generate and spread disinformation.