
Towards Unified Multimodal Interleaved Generation via Group Relative Policy Optimization

Ming Nie¹

Chunwei Wang²

Jianhua Han²

Hang Xu²

Li Zhang¹

¹School of Data Science, Fudan University

²Noah’s Ark Lab, Huawei

A Data Preparation

During the warm-up stage, we collect 0.3M text-image paired samples with interleaved outputs from ActivityNet [1], GenHowTo [3] and OpenStory++ [4]. 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 [2]. 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.

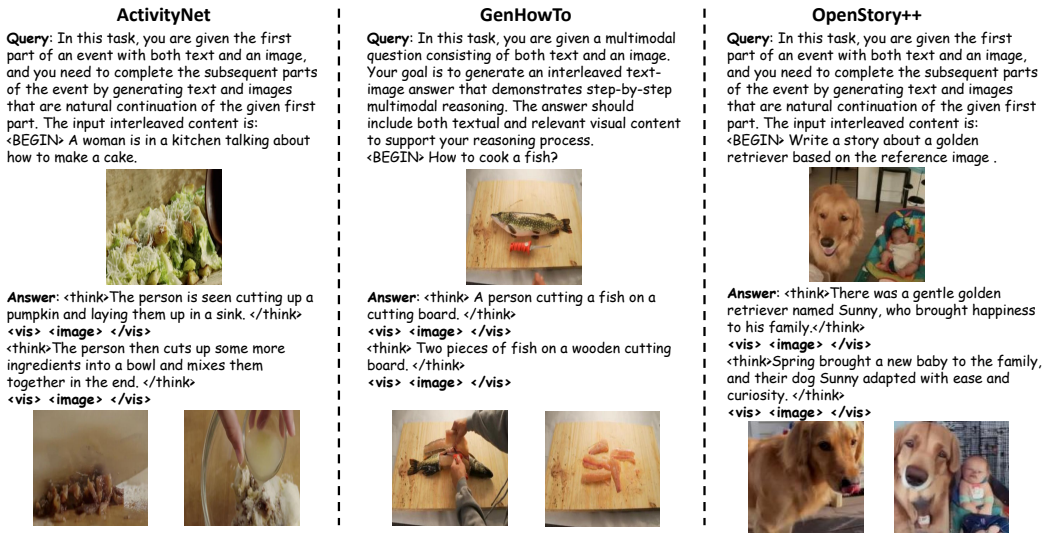


Figure 1: Illustration of data preparation process during the warm-up stage.

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

Input: Question: How to Stop Social Media from Damaging Your Offline Behavior?
Method 1: Balancing Online and Offline Social Networks
Step 1: Make offline networking a priority



Output: <think>Step 2: Set boundaries for social media usage. Designate specific times of day to check social media so it doesn't interfere with real-life interactions.</think>
<vis> <image> </vis>



Figure 2: Failure cases analysis about hallucinated visual content.

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 1.

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 1.

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 2. First, in some complex reasoning scenarios, the model may generate hallucinated visual content that does not 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.

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.
- [2] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.
- [3] 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.
- [4] 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.