Causal Visual Narration: Modeling Event Causality in Image Sequence Storytelling

Anonymous EMNLP 2026 submission

Abstract

Image-sequence visual storytelling requires multimodal large language models (MLLMs) to track entity states, event progression, and causal relations across ordered images. However, many existing MLLM-based methods encode each image independently and rely on the language model or high-level semantic modules to fuse visual tokens, which may fail to preserve cross-frame dependencies. As a result, generated stories often degenerate into disconnected static descriptions with inconsistent characters, discontinuous events, and weak causal coherence. To address this problem, we propose **Causal Visual Narration (CVN)**, a causal-aware framework that models change-driven dependencies for multi-image storytelling. CVN first uses a difference-aware Causal Q-Former to inject cross-frame semantic changes into visual tokens. It then constructs narrative scene graphs for the image sequence, infers inter-image causal bridges from cross-frame changes, and aggregates this information into a causal narrative card to guide story generation. Experimental results demonstrate that CVN achieves superior performance over existing models, producing image-sequence narratives with stronger visual faithfulness, event coherence, and narrative completeness. Our code is available at: https://anonymous.4open.science/r/CNV-5403.

1 Introduction

Multimodal large language models (MLLMs) have made rapid progress by connecting visual encoders with language models through projection layers, Q-Formers, or cross-attention modules (Alayrac et al., 2022; Tsimpoukelli et al., 2021; Li et al., 2023b; Dai et al., 2023; Liu et al., 2023; Zhu et al., 2023; Radford et al., 2021). Yet image-sequence visual storytelling goes beyond multi-image captioning, requiring entity tracking and event-transition reasoning across an ordered image sequence to produce a causally coherent narrative. (Huang et al., 2016).

This requirement exposes a weakness of common multi-image MLLMs. Many systems encode each image separately and perform fusion only in a high-level semantic or language context (Li et al., 2023a; Zhao et al., 2023; Wu et al., 2024). Such late fusion can recognize salient objects in each frame, but it often loses evidence about what has changed across frames. Consequently, generated stories may degenerate into concatenations of static captions, confuse character identities, or introduce causally disconnected events (Kim et al., 2018; Wang et al., 2024; Park et al., 2020).

Prior visual-storytelling methods improve decoding with attention, reinforcement learning, external knowledge, scene graphs, or preference optimization (Wang et al., 2018; Kim et al., 2018; Hsu et al., 2020; Yang et al., 2025). These methods are useful, but the core cue in our setting is often not a single object or a generic knowledge link; it is a cross-image change that explains why a later state follows from earlier evidence. Figure 1 gives a motivating example: independent visual descriptions miss the narrative bridge, while change-aware causal modeling makes the event relation explicit. This motivates a shift from static visual recognition to change-driven event modeling for causally coherent storytelling.

We propose Causal Visual Narration (CVN), a causal modeling framework designed to turn cross-frame changes into explicit narrative evidence for image-sequence storytelling. To address the loss of transition information in late fusion, CVN first introduces a Causal Q-Former that compares adjacent images and incorporates their semantic differences into visual-token representations. This enables the model to emphasize evolving entities, states, actions, and interactions, rather than treating each image as an isolated static scene. CVN further constructs narrative scene graphs from the image

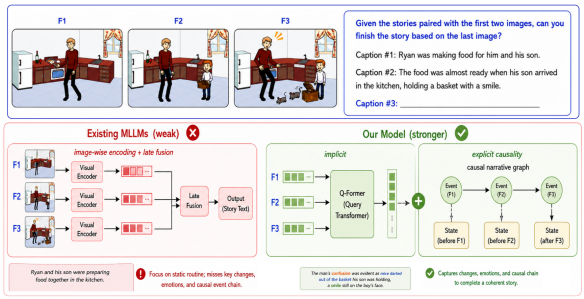

Figure 1: Motivating example. Existing models may describe separate images but miss the cross-frame event link. CVN extracts semantic changes and uses them as causal evidence for story generation.

sequence and caption context, identifies dominant cross-frame changes, and infers a compact latent causal bridge that explains how earlier visual evidence leads to later narrative states. The inferred bridge and its supporting evidence are organized into a causal narrative card, which records key participants, transition cues, grounding evidence, and verifiable narrative elements. During generation, this card provides lightweight planning and verification guidance, helping the model follow the intended event direction and revise the story only when necessary. In this way, CVN directly targets the static-caption, identity-confusion, and disconnected-event errors introduced above.

Our contributions are summarized as follows:

- We propose Causal Visual Narration (CVN), a causality-driven framework for image-sequence storytelling that transforms cross-frame visual changes into explicit narrative evidence, enabling story generation to better model event evolution beyond static image-level semantics.
- We develop two complementary causal modeling components in CVN: a Causal Q-Former that learns difference-aware visual tokens from adjacent images, and a causal narrative card that organizes dominant transitions, subject entities, grounding evidence, and latent causal bridges for planning and verification during generation.
- We conduct experiments on VIST and AESOP. Results from automatic evaluation, ablation studies, and qualitative case analyses demonstrate the effectiveness of CVN in improving narrative coherence, character consistency, and cross-event logical continuity.

2 Related Work

2.1 Visual Storytelling

Visual storytelling generates coherent narratives from ordered image sequences and requires modeling entity continuity and event progression (Huang et al., 2016). Early methods improved story generation with recurrent encoder–decoder architectures, attention mechanisms, or reinforcement learning (Wang et al., 2018; Kim et al., 2018), while recent work incorporates pretrained language models, external knowledge, scene-level relations, visual commonsense, or MLLM-based adaptation to enrich narrative content (Hsu et al., 2020; Speer et al., 2017; Krishna et al., 2017; Zellers et al., 2019; Park et al., 2020; Bosselut et al., 2019; Wang et al., 2024; Yang et al., 2025). These approaches improve semantic richness and generation quality, but they mainly operate through decoding, knowledge augmentation, or model adaptation, leaving cross-frame visual changes insufficiently modeled before generation. This motivates CVN to use such changes as narrative evidence for image-sequence storytelling.

2.2 Multimodal Large Language Models

Recent MLLMs have rapidly advanced from single-image understanding toward multi-image and interleaved visual reasoning. Representative open-source models such as BLIP-2, InstructBLIP, and LLaVA establish strong visual-language alignment through Q-Formers, instruction tuning, or visual projectors (Li et al., 2023b; Dai et al., 2023; Liu et al., 2023), while frontier systems such as GPT-4o and Gemini further demonstrate broad multimodal interaction capabilities (OpenAI, 2024; Gemini Team and Google DeepMind, 2025). More recent multi-image MLLMs, including Qwen2.5-VL, InternVL3, LLaVA-NeXT-Interleave, and Mantis, extend this direction by supporting dynamic visual inputs, interleaved image-text contexts, or multi-image instruction tuning (Qwen Team, 2025; Zhu et al., 2025; Li et al., 2024; Jiang et al., 2024). Despite these advances, most multi-image MLLMs still fuse independently extracted visual tokens through interleaved context modeling or language-level attention. This limits explicit modeling of cross-frame changes, motivating CVN to treat such changes as causal evidence for image-sequence storytelling.

2.3 Causal Narrative Modeling

Causal reasoning is fundamental to narrative understanding, as stories should explain how earlier events lead to later states rather than merely arrange events in temporal order. In language, this problem has been studied through narrative event chains, story-cloze evaluation, event-centered commonsense inference, and event causality identification (Chambers and Jurafsky, 2008; Mostafazadeh et al., 2016; Rashkin et al., 2018; Sap et al., 2019; Bosselut et al., 2019; Do et al., 2011). In vision-language settings, causal relations are harder to model because the relevant evidence is often implicit and may only appear as changes across frames. This is especially challenging for image-sequence storytelling, where sparse images may omit key bridge events. Existing visual commonsense, scene-graph, and knowledge-enhanced methods can model objects, spatial relations, intentions, or likely consequences (Krishna et al., 2017; Johnson et al., 2018; Hsu et al., 2020; Speer et al., 2017; Zellers et al., 2019; Park et al., 2020), but they are not primarily designed to derive event-level causal bridges from cross-frame visual changes. Motivated by causal representation learning (Scholkopf et al., 2021), CVN treats such changes as evidence for latent narrative causality in image-sequence storytelling.

3 Task Formulation

Given an ordered image sequence $\mathcal{X} = \{I_t\}_{t=1}^{T}$, the goal is to generate a story $\mathcal{Y} = \{y_1, \ldots, y_L\}$ that is visually grounded, temporally coherent, and causally plausible (Huang et al., 2016; Chambers and Jurafsky, 2008; Mostafazadeh et al., 2016; Bhagavatula et al., 2020). We represent the underlying narrative as

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B}),$$

where $\mathcal{V}$ contains observed entities, actions, and states; $\mathcal{E}$ contains cross-image relations such as entity continuity, state transition, and action progression; and $\mathcal{B}$ contains latent bridge events needed to explain unobserved transitions. A bridge $b = (\mathcal{V}_{\leq t}, r, \mathcal{V}_{>t}, c)$ connects earlier evidence to a later state through an event description r and optional commonsense support c.

4 Method

CVN is organized around two complementary stages: an implicit stage that enhances visual representations with cross-frame transition cues, and an explicit stage that structures these cues into causal constraints for story generation.

4.1 Overview

Given an ordered image sequence $\mathcal{X} = \{I_1, I_2, \ldots, I_T\}$, a shared vision encoder first maps each image into patch-level semantic features

$$F_t = E_v(I_t) = \{f_{t,j}\}_{j=1}^{P}, \quad t = 1, \ldots, T,$$

where P denotes the number of visual patches, following patch-based visual representation learning (Dosovitskiy et al., 2021; Radford et al., 2021). The Causal Q-Former then compares adjacent image features in the same semantic space and produces a causality-enhanced visual representation Z_t (Figure 2). The multimodal causal graph module further constructs story-level causal dependencies from scene graphs, cross-frame changes, and latent bridge events (Figure 3). Finally, the generation module transforms the causal graph into executable constraints, including bridge prompts and verifiable narrative units, to guide the language model toward causally coherent story endings.

4.2 Module 1: Causal Q-Former

CVN introduces a difference-aware Causal Q-Former to capture narrative-relevant changes during visual-token construction. Inspired by query-based visual-language alignment and temporal representation learning (Li et al., 2023b; Bertasius et al., 2021), the module compares adjacent image features in a shared semantic space and incorporates their semantic differences into the current-frame representation. In this way, the model can emphasize evolving entities, states, actions, and interactions before story generation.

Semantic difference modeling. Given an ordered image sequence, each frame is first encoded by a shared vision encoder $F_t = E_v(I_t)$. For the current frame I_t, CVN summarizes its local semantic evolution by fusing backward and forward differences:

$$\bar{\Delta}_t = \mathrm{MLP}_{\Delta}([F_t - F_{t-1}; F_{t+1} - F_t]),$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation. The backward difference captures how the current state emerges from the previous frame, while the forward difference reflects how it relates to the following frame. The fused representation $\bar{\Delta}_t$ is used as causal context rather than a replacement for the current-frame feature F_t. For boundary frames, the missing neighboring feature is replaced by a zero vector or the available one-sided difference.

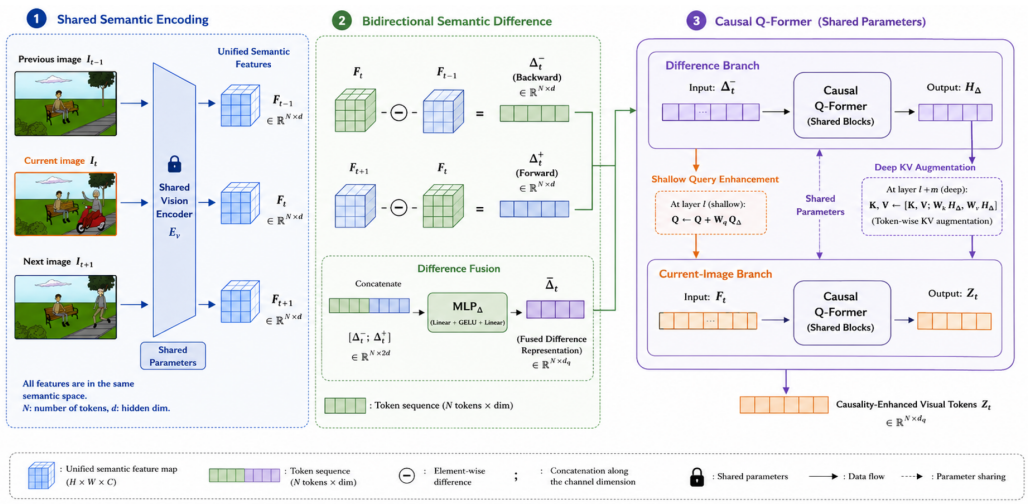

Figure 2: Causal Q-Former. CVN injects semantic differences from adjacent frames into current-image visual tokens through query enhancement and key–value augmentation.

Difference-aware token interaction. The Causal Q-Former contains a difference branch and a current-image branch with shared structure. The difference branch encodes the fused semantic difference $\bar{\Delta}_t$ and produces a change-aware hidden representation H_t^Δ, while the current-image branch encodes the current-frame feature F_t. The goal is to make the current-frame visual tokens attend not only to static image content, but also to cross-frame semantic changes.

At a shallow layer l, the query representation of the current-image branch is enhanced by the difference branch:

$$Q_t^{I,l} \leftarrow Q_t^{I,l} + W_q Q_t^{\Delta,l},$$

where $Q_t^{I,l}$ and $Q_t^{\Delta,l}$ denote the query representations of the current-image branch and the difference branch, respectively. This operation guides the current-image branch to focus on regions related to cross-frame changes.

At a deeper layer $l+m$, the change-aware hidden representation is injected into the key–value states of the current-image branch, following the query–key–value attention formulation of Transformers (Vaswani et al., 2017). For the key states, this operation is written as

$$K_t^{I,l+m} \leftarrow [K_t^{I,l+m}; W_k H_t^\Delta],$$

where H_t^Δ denotes the hidden representation from the difference branch and W_k is a learnable projection. The value states are updated in the same way with an independent projection matrix. Through this key–value augmentation, the current-image branch can access cross-frame change information during attention aggregation. After these interactions, the current-image branch continues to the final Q-Former layer. Its final hidden states are used as the causality-enhanced visual tokens: $Z_t = H_t^{I,L}$, where L denotes the last Q-Former layer, while l and $l+m$ denote the intermediate interaction layers. The token sequence Z_t is then passed to the language model for story generation.

4.3 Module 2: Causal Narrative Card and Guided Generation

The difference-aware visual tokens produced by Module 1 encode cross-frame changes implicitly. However, story generation also requires an explicit structure that specifies which changes are narratively important and how they support the final event. Module 2 addresses this need by constructing a causal narrative card, which converts visual evidence and caption context into structured constraints for generation. The card is designed to bridge implicit change perception and explicit narrative planning.

Narrative scene graph construction. Following scene-graph representations for structured visual understanding (Krishna et al., 2017; Xu et al., 2017; Johnson et al., 2018), given an ordered image sequence $I = \{I_1, I_2, \ldots, I_T\}$, each frame is represented as a narrative scene graph

$$G_t = (O_t, A_t, S_t, R_t),$$

where O_t, A_t, S_t, and R_t denote visible entities, actions, states, and spatial relations, respectively. The scene graph provides visually grounded evidence for later reasoning, but it does not directly encode

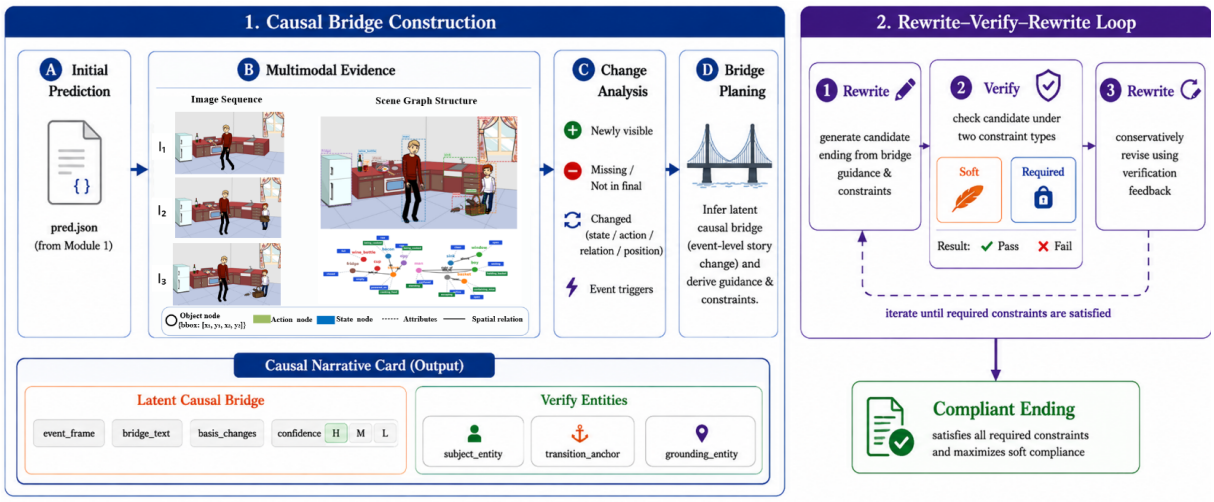

Figure 3: Overview of causal narrative card construction. CVN organizes scene-graph evidence, cross-frame changes, and latent causal bridges into verifiable constraints for story generation.

hidden intentions or causal conclusions. This separation ensures that causal bridges are inferred from cross-frame changes rather than from unsupported single-frame assumptions.

Causal bridge reasoning. CVN compares adjacent scene graphs together with caption context to identify narrative-relevant changes, including entity changes, action or state transitions, relation shifts, and caption-triggered event cues. Stable entities are retained as contextual anchors, while dominant changes are used as evidence for event progression.

Formally, the change set between two adjacent frames is defined as

$$\Delta_t = \phi(G_t, G_{t+1}, C),$$

where C denotes the caption context. Based on this change set, CVN infers a compact latent causal bridge:

$$Z_t = f_\theta(G_t, G_{t+1}, C, \Delta_t).$$

Here, Z_t represents the event-level transition that explains how the story moves from one visual state to the next, similar in spirit to abductive commonsense reasoning over incomplete event observations (Bhagavatula et al., 2020). Instead of expanding all possible causal hypotheses, CVN keeps a sparse bridge so that generation is guided by the most salient narrative change.

Causal narrative card construction. The inferred bridge and its supporting evidence are organized into a causal narrative card:

$$C = (F^{obs}, Z, U^{ver}, K^{cs}, S^{gen}),$$

where F^{obs} denotes observed scene-graph facts, Z is the latent causal bridge, U^{ver} contains verifiable narrative units, K^{cs} represents commonsense support, and S^{gen} denotes generation constraints. The verifiable units are not a complete list of visible objects; they correspond to event-functional elements such as key participants, transition anchors, grounding cues, and outcome states. In this way, the card prioritizes changes that are important for narrative causality rather than visually salient but irrelevant background details.

Card-guided rewriting After constructing the causal narrative card, CVN uses it to refine the initial story ending only when necessary. Let y^0 denote the initial ending and C denote the causal narrative card. The refinement process is defined as

$$\hat{y} = \begin{cases} y^0, & \text{Verify}(y^0, C) = 1, \\ \text{Rewrite}(y^0, C), & \text{Verify}(y^0, C) = 0. \end{cases}$$

The verification function checks whether the candidate ending expresses the latent bridge and covers the required verifiable units. This check is semantic rather than purely lexical, allowing paraphrases or equivalent event expressions when they preserve the intended causal relation, following the broader idea of generation-time self-refinement and verification (Madaan et al., 2023).

If verification fails, CVN performs a conservative rewrite guided by the causal card. The latent bridge specifies the event direction, while the verifiable units constrain the key participants, transition cues, and outcome states that should be preserved. The rewrite therefore aims to repair missing or incorrect causal relations without replacing useful content from the initial prediction. This design allows CVN to correct final-frame-only or weakly grounded

endings while reducing the risk of uncontrolled rewriting.

4.4 Training

For efficient training, we freeze the vision encoder, LLM, and original Q-Former layers, and only update the visual-language projection layer and the newly introduced causal modules in Module 1, resulting in 6.1M trainable parameters, which account for about 0.09% of the 7B backbone. The model is optimized with the autoregressive language modeling objective:

$$\mathcal{L} = -\sum_{i=1}^{t} \log P(T_i \mid V, T_{<i}; \theta),$$

where V denotes the causality-enhanced visual tokens, $T_{<i}$ denotes the preceding textual tokens, T_i is the current target token, and θ represents the trainable parameters. Training is performed in two stages: we first adapt the visual-language projection layer on MmLINK and then optimize the causal modules while keeping the projection layer frozen. Both stages use AdamW (Loshchilov and Hutter, 2019) with a learning rate of 2×10^{-5}, weight decay of 0.05, and linear warmup cosine decay. Training is conducted for 2 epochs on 4 A800 GPUs with mixed precision and a per-GPU batch size of 20.

5 Experiments

5.1 Experimental Settings

We adopt InstructBLIP-Vicuna-7B (Dai et al., 2023) as the base model and resize all images to 224×224. Story endings are decoded by beam search with `num_beams=4`. For causal-card refinement, Qwen2.5-7B (Qwen Team, 2025) is used as the planner with low-temperature decoding, followed by local verification and conservative rewriting.

Datasets. We evaluate CVN on two image-sequence storytelling datasets: VIST (Huang et al., 2016) and AESOP (Ravi et al., 2021). VIST contains ordered photo albums with human-written stories, while AESOP focuses more on story-level event progression. Following common practice in visual storytelling (Wang et al., 2018; Kim et al., 2018; Yang et al., 2025), we compare generated stories with reference stories using automatic text metrics.

Baselines. We compare CVN with representative MLLMs and visual storytelling methods, including BLIP-2 (Li et al., 2023b), InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023), GPT-4o-mini (OpenAI, 2024), Gemini 2.5 (Gemini Team and Google DeepMind, 2025), Qwen2.5-VL (Qwen Team, 2025), InternVL2 (Chen et al., 2024), and SAM (Wu et al., 2024). Among them, SAM is the most related baseline because it also targets cross-image token alignment, allowing us to examine whether causal bridge modeling brings additional gains beyond semantic alignment.

Metrics. We report ROUGE-L, CIDEr, BLEU-4, and METEOR (Lin, 2004; Vedantam et al., 2015; Papineni et al., 2002; Banerjee and Lavie, 2005). These metrics are standard in visual storytelling and mainly measure reference overlap and content consistency. Since they cannot fully reflect causal coherence, we further include ablation studies and qualitative analysis.

5.2 Performance Comparison

The overall results of CVN and all baselines are reported in Table 1. CVN achieves the best performance on both VIST and AESOP across all automatic metrics, demonstrating the effectiveness of causal change modeling for image-sequence visual storytelling. Compared with the most related SAM baseline, CVN brings consistent improvements on both datasets, with the most pronounced gain observed on CIDEr.

1.CVN enhances the ability to model cross-frame event progression. Many MLLM baselines can recognize salient content in individual images, but they often fail to connect visual changes into a coherent story transition. By injecting semantic differences into visual-token construction and further organizing them as latent causal bridges, CVN better captures how entities, states, and events evolve across the sequence. This explains its large gains on content-sensitive metrics such as CIDEr and ROUGE-L.

2.CVN improves narrative grounding and causal closure. Existing multi-image methods tend to rely on target-frame information or generic narrative patterns, which may lead to static caption concatenation, identity inconsistency, or weak event resolution. In contrast, CVN uses verifiable narrative units to preserve bridge-critical participants, transition cues, and outcome states during generation. The consistent improvements over strong baselines such as Qwen2.5-VL-7B and SAM suggest that causal bridge modeling provides benefits beyond general multi-image perception or semantic

Model	VIST				AESOP			
	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR
InstructBLIP	12.61	18.00	1.24	4.98	14.14	6.87	0.20	5.39
BLIP-2	13.41	7.26	0.69	7.72	17.96	8.88	2.18	12.67
LLaVA	14.80	19.55	1.49	8.00	17.89	17.74	2.01	8.98
GPT-4o-mini	11.22	2.94	0.64	8.09	17.23	17.11	2.02	10.28
Gemini 2.5	13.84	22.10	1.74	7.13	20.52	21.50	3.37	9.72
Qwen2.5-VL-7B	20.08	43.62	3.41	9.69	23.11	34.12	4.03	11.92
InternVL2-8B	12.94	13.24	1.20	7.27	19.55	22.22	2.47	9.96
SAM	20.85	39.32	3.65	8.50	23.45	25.92	4.13	11.81
CVN (ours)	**26.95**	**66.01**	**5.63**	**12.53**	**24.77**	**34.63**	**4.50**	**13.43**

Table 1: Main results on VIST and AESOP. Best scores are shown in bold.

alignment alone.

5.3 Component Analysis

The ablation results show that the main components of CVN are complementary and jointly contribute to the final performance.

1.difference-aware visual encoding is the most critical source of improvement. Row 2 substantially outperforms Row 1, especially on CIDEr and BLEU-4, showing that cross-frame semantic differences should be captured during visual-token construction rather than inferred only after generation. In contrast, Row 3 brings little benefit, indicating that generic rewriting alone cannot improve story generation without reliable causal evidence.

2.latent causal bridge reasoning is necessary for turning visual evidence into narrative-relevant event structure. Although narrative scene graphs provide grounded observations, Row 4 performs worse than Row 2, suggesting that scene-

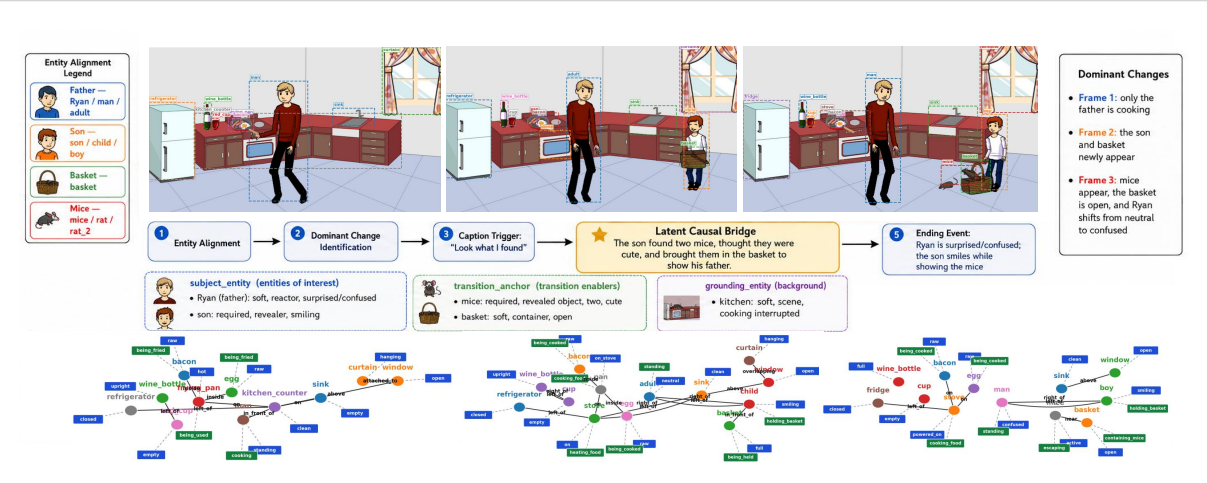

Figure 4: Case study of causal bridge construction. CVN transforms cross-frame visual changes into a latent event bridge and verifiable narrative constraints.

Context Captions
#1 We attended an awesome New Year's Eve party.
#2 The drinks were flowing fast and furious.
#3 The two DJs kept us dancing until the big event.
#4 Finally it was time! The place exploded in celebration.
Task: Generate the story climax based on the target image.

Model	Generated Ending
Human Ref.	Happy New Year!
SAM	Everyone was having a great time.
Qwen2.5-VL-7B	The party was a huge success and we had a great time!
LLaVA	The countdown to midnight was intense.
BLIP2	The DJs kept us dancing until the big event.
GPT-4o	As the night wound down, we left the party buzzing with excitement, ready to embrace the new year ahead.
CVN (Ours)	**Happy New Year to all!**

Context Captions
#1 Ryan was making food for him and his son.
#2 The food was almost ready when his son arrived in the kitchen. "Look what I found," said the son.
Task: Complete the story based on the target image.

Model	Generated Ending
Human Ref.	He had found two mice, found them cute and brought them to show to his father.
SAM	Ryan and his son were preparing food together in the kitchen.
Qwen2.5-VL-7B	The son had found two mice in the kitchen.
LLaVA	The son brought a basket of food to the kitchen.
BLIP2	Ryan was making food for him and his son. The food was almost ready when his son arrived in the kitchen.
GPT-4o	Look, Dad! I found a little mouse in the garden, can we keep it? his son exclaimed, holding the basket with excitement.
CVN (Ours)	**The man's confusion was evident as mice darted out of the basket his son was holding, a smile still on the boy's face.**

Figure 5: MLLM comparison on story climax prediction. CVN produces more causally coherent endings with cross-frame evidence.

	$+\Delta_{CQF}$	$+\Delta_{RW}$	$+\Delta_{NSG}$	$+\Delta_{LCB}$	$+\Delta_{VNU}$	$+\Delta_{VER}$	Average			
							ROUGE-L	CIDEr	BLEU-4	METEOR
1							15.68	8.07	1.44	10.20
2	√						22.65	33.22	4.35	10.46
3		√					15.79	8.11	1.17	10.67
4	√	√	√				18.09	18.66	2.63	10.62
5	√	√		√			23.15	34.13	3.89	10.66
6	√	√		√	√		24.08	39.11	3.12	11.03
7 (ours)	√	√		√	√	√	**25.86**	**50.32**	**5.07**	**12.98**

Table 2: Ablation study of causal modules in CVN, averaged over VIST and AESOP. $+\Delta_{CQF}$: difference-aware Causal Q-Former; $+\Delta_{RW}$: rewriting; $+\Delta_{NSG}$: narrative scene graphs; $+\Delta_{LCB}$: latent causal bridge; $+\Delta_{VNU}$: verifiable narrative units; $+\Delta_{VER}$: verification and repair.

graph constraints alone may introduce irrelevant visual details. After adding latent bridge reasoning in Row 5, performance recovers substantially, which shows that the bridge module helps filter raw visual evidence and organize dominant changes into an event-level explanation.

3.verifiable narrative constraints and conservative verification further improve the controllability of generation. Row 6 improves over Row 5 on most metrics, indicating that verifiable narrative units help preserve bridge-critical participants, transition anchors, and outcome states. The full model in Row 7 achieves the best results across all metrics, showing that verification-based repair can effectively control the rewriting process and prevent causal constraints from becoming uncontrolled generation noise.

5.4 Human Evaluation Results

To further assess story quality beyond automatic metrics, we conduct a small-scale human evaluation on 5 randomly sampled examples from each dataset. We compare CVN with three representative baselines and anonymize all generated stories before evaluation. Three annotators rate each story on a 1–5 scale in terms of visual relevance, attractiveness, and coherence, and we report their geometric mean (GM) as the overall score. As shown in Table 3, CVN achieves the best scores in visual relevance, coherence, and overall GM, demonstrating its stronger human-perceived story quality.

Method	Rel.	Attr.	Coh.	GM
SAM	3.10	3.25	2.85	3.06
Qwen2.5-VL-7B	4.05	3.70	3.75	3.83
GPT-4o	3.55	**4.20**	3.45	3.72
CVN (Ours)	**4.60**	4.05	**4.55**	**4.39**

Table 3: Case-level human evaluation. Rel., Attr., Coh., and GM denote visual relevance, attractiveness, coherence, and geometric mean.

5.5 Case Study

Figure 4 illustrates how CVN converts cross-frame visual evidence into an explicit causal planning signal. Instead of relying on the final frame alone, CVN aligns recurring entities, identifies dominant changes across the sequence, and infers a latent bridge that explains the transition between visual states. The bridge is then mapped into verifiable narrative units, which constrain the generated ending to preserve key participants, transition cues, and event outcomes. This case shows that CVN helps reduce final-frame-only descriptions and supports story generation with clearer event progression and causal coherence.

Figure 5 compares different vision-language models in story climax generation. Existing methods often produce fluent but weakly grounded endings, because they tend to rely on the target frame or generic narrative patterns rather than the event transition implied by the full sequence. CVN addresses this issue by using cross-frame changes to infer a latent causal bridge and by constraining generation with verifiable narrative units, leading to endings with clearer event progression, better character consistency, and stronger causal closure.

6 Conclusion

In this work, we introduced CVN, a causal-aware framework for image-sequence visual storytelling that emphasizes cross-frame transitions rather than isolated frame descriptions. By combining difference-aware visual encoding with explicit causal bridge guidance, CVN improves event continuity and causal coherence in generated stories. Our results further show that causal constraints are most effective when they are visually grounded and applied conservatively, suggesting that future work should focus on confidence-aware bridge planning and causal-coherence evaluation beyond surface text overlap.

Limitations

CVN relies on intermediate structured evidence, including scene graphs, entity alignment, cross-frame change detection, and latent bridge inference. Errors in these components may propagate to the causal narrative card and lead to inaccurate or overly restrictive generation constraints, especially when images contain occlusion, ambiguous interactions, or multiple plausible event interpretations. In addition, the current framework focuses on compact causal bridges derived from dominant cross-frame changes. While this design improves controllability, it may be less effective for complex stories involving long-range dependencies, multiple simultaneous events, or competing causal explanations. Future work will explore confidence-aware causal planning and richer event-chain modeling for more complex narrative scenarios.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of ACL*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of EMNLP*, pages 294–303.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Gemini Team and Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind.

Chao-Chun Hsu, Zhi-Yuan Chen, Chih-Yang Hsu, Chih-Chia Li, Ting-Yao Lin, Ting-Hao Huang, and Lun-Wei Ku. 2020. Knowledge-enriched visual storytelling. In *Proceedings of AAAI*.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of NAACL-HLT*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. MANTIS: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228.

Taehyeong Kim, Moonsu Heo, Seonil Son, Kyoung-Woon Park, and Byoung-Tak Zhang. 2018. GLAC Net: GLocal attention cascading networks for multi-image cued story generation. *arXiv preprint arXiv:1805.10973*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. LLaVA-NeXT-Interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of ICML*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. Technical report, OpenAI.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. VisualCOMET: Reasoning about the dynamic context of a still image. In *Proceedings of ECCV*.

Qwen Team. 2025. Qwen2.5-VL technical report. Technical report, Qwen. ArXiv preprint arXiv:2502.13923.

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

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of ACL*, pages 463–473.

Hareesh Ravi, Kushal Kafle, Scott Cohen, Jonathan Brandt, and Mubbasir Kapadia. 2021. AESOP: Abstract encoding of stories, objects, and pictures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of AAAI*, pages 3027–3035.

Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Eileen Wang, Chia-Chun Han, and Josiah Poon. 2024. SCO-VIST: Social interaction commonsense knowledge-based visual storytelling. In *Proceedings of EACL*, pages 1602–1616.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of ACL*.

Tao Wu, Ming Li, Jun Chen, Wei Ji, Weiyao Lin, Junyu Gao, Kun Kuang, Zhou Zhao, and Fei Wu. 2024. Semantic alignment for multimodal large language models. In *Proceedings of ACM Multimedia*.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419.

L. Yang, Z. Xiao, W. Huang, and X. Zhong. 2025. StoryLLaVA: Enhancing visual storytelling with multimodal large language models. In *Proceedings of COLING*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of CVPR*.

Haoyu Zhao, Zefan Cai, Siqi Si, Xiaojian Ma, Kai An, Liang Chen, Zewen Liu, Sheng Wang, Wenxuan Han, and Baobao Chang. 2023. MMICL: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuxuan Duan, et al. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Supplementary Material

A Additional Ablation Protocols

This section provides additional diagnostic protocols for analyzing the causal bridge planner and the rewriting stage. These protocols complement the main ablation study in Table 2 and are intended to clarify which evidence sources and runtime constraints contribute to the final generation quality.

A.1 Causal Bridge Extraction Variants

Table 4 lists diagnostic variants for analyzing the bridge planner. These variants are not intended to replace the main ablation in Table 2; instead, they explain which evidence sources and prompt rules are responsible for bridge quality.

A.2 Rewrite and Verification Variants

Table 5 isolates the rewrite stage implemented in `rewrite_from_pred_json.py`. All rows use the same first-module `pred.json` and the same planner payload unless otherwise stated.

Interpretation. Three patterns are worth emphasizing. First, *Base SAM* is the strongest variant under automatic metrics on both datasets, which indicates that the first-module prediction already captures much reference-overlapping surface content. Second, rewrite-based variants—especially *Scene-graph constraints*, *Bridge + required units*, and *Full CVN*—often reduce ROUGE-L and CIDEr on VIST, where endings are short and highly conventional; forcing event-level bridges can replace a partially correct sentence with a causally focused but lexically divergent one. Third, on AESOP the degradation is smaller and *Latent bridge only* remains competitive, suggesting that event direction helps when stories contain clearer action progression. Together with the case study in Section 5.5, these results support a conservative deployment strategy: apply full verification and rewriting only when bridge confidence is high, and fall back to the initial prediction when constraints are uncertain or surface-only.

B Human Evaluation Protocol

For the human evaluation in Section 5.4, annotators are shown the ordered image sequence, the preceding captions, and an anonymized generated ending. Model names are hidden during annotation. Annotators rate each ending on a 1–5 scale according to three criteria: visual relevance, attractiveness, and coherence. Visual relevance measures whether the ending is consistent with the visual content and avoids hallucinated information. Attractiveness measures whether the ending is fluent, natural, and engaging. Coherence measures whether the ending logically continues the preceding story and forms a coherent event progression. The annotation form is shown in Table 6.

C Prompt Templates

This section gives paper-formatted versions of the prompts used in the generation pipeline. Dataset-specific image ids, captions, scene graphs, and candidate endings are replaced with placeholders.

Variant	Removed or restricted information	Diagnostic purpose
Caption-only bridge	Remove scene graphs from the planner input	Tests whether story captions alone produce fluent but visually ungrounded endings.
Final-frame-only bridge	Keep only the final-frame scene graph	Measures final-image bias and the tendency to describe visible objects instead of story resolution.
Cross-frame graph bridge	Use scene graphs without caption-trigger rules	Tests whether visual changes alone recover the intended narrative event.
No missing-entity rule	Do not prioritize entities marked as missing in the final frame	Tests loss, search, theft, disappearance, and ownership-transfer cases.
No event-prior rules	Remove case-derived rules for holidays, accidents, pranks, rejection, competitions, and searches	Tests whether event priors improve short or conventional VIST endings.
Full planner	Caption context plus cross-frame scene graphs and event-prior rules	Full bridge extraction used by CVN.

Table 4: Diagnostic ablations for causal bridge extraction. The expected evaluation includes automatic metrics and bridge-level human or LLM judgment of whether the inferred bridge matches the reference ending event.

Variant	Runtime setting	Diagnostic purpose
No rewrite	Use `-no-rewrite` or export `pred_response_initial`	Measures the first-module baseline before causal intervention.
Unconstrained rewrite	Remove `bridge_context` and `story_constraints` from the rewrite prompt	Controls for generic LLM paraphrasing.
Bridge-prompt rewrite	Keep only `latent_causal_bridge.bridge_text`	Tests whether event direction alone is enough.
Verify-only selection	Run local verification but do not call the rewrite model	Tests whether the initial output already satisfies causal slots.
Generate–verify–rewrite	Enable local verification and bounded rewrite attempts	Tests the executable constraint loop.
Full repair with overlap guard	Enable repair and keep `min_rewrite_overlap`	Tests whether conservative repair prevents harmful over-rewriting.

Table 5: Diagnostic ablations for the rewrite and verification stage. Besides automatic metrics, we report required-unit coverage and rewrite harm rate.

Field	Description
Sample ID	The anonymized id of the evaluated image sequence.
Image Sequence	The ordered image sequence shown to annotators.
Context Captions	The preceding story captions provided as narrative context.
Generated Ending	The anonymized model output to be evaluated.
Visual Relevance	Score from 1 to 5. Whether the ending is consistent with the visual content and avoids hallucinated information.
Attractiveness	Score from 1 to 5. Whether the ending is fluent, natural, and engaging.
Coherence	Score from 1 to 5. Whether the ending logically continues the preceding story and forms a coherent event progression.
Comment	Optional free-form explanation of major errors or preference.

Table 6: Human evaluation annotation form.

Scene-Graph Generation Prompt
Instruction: Analyze the image and construct a visually grounded scene graph. Identify all visible entities, assign each entity a bounding box, one action node, and one state node, and extract only spatial relationships between entities.
Input: A single image.
Entity schema: Each object node contains `entity`, `bbox`, `Action Nodes`, and `State Nodes`. Action nodes describe what the entity is doing or undergoing; state nodes describe visible condition, emotion, or result state.
Relation schema: Spatial relationships use exactly `subject`, `relation`, and `object`. Allowed relations include `left_of`, `right_of`, `above`, `below`, `in_front_of`, `behind`, `overlapping`, `inside`, `intersecting`, `attached_to`, `on`, and `under`.
Constraints: Do not infer causal relationships, hidden intentions, or unobserved events. Do not output reasoning or analysis. Use concise labels and prefer `snake_case`.
Output format: Return only one `<scene_graph>` block containing JSON with `Object Nodes` and `Spatial Relationships`.

Table 7: Prompt template for scene-graph generation.

Causal Bridge Planner Prompt
Instruction: Given story captions and consecutive scene graph JSONs, infer one short event-level latent causal bridge and extract the key verify entities for the final sentence. The goal is not to describe the final image alone; the final sentence should resolve the caption setup.
Input: Sample id; story captions/instruction context; compact scene graphs for all frames.
Reasoning tasks: Align caption entities with scene-graph entities; find dominant story changes, including newly visible entities, missing entities, action/state shifts, relation changes, and caption event triggers; infer one bridge from dominant changes plus commonsense.
Bridge constraints: Prefer story-level goals, conflicts, searches, losses, accidents, pranks, celebrations, reactions, or outcomes over literal pose descriptions. Stable entities should usually serve as witnesses or grounding context, not the bridge center. Missing prior entities or objects can be central to the bridge.
Output schema: Return a JSON object with two top-level fields: `latent_causal_bridge` and `verify_entities`. The bridge field contains `event_frame`, `bridge_text`, `basis_changes`, and `confidence`; `event_frame` is selected from search, questioning, loss/theft, concern, pursuit, interruption, reaction, continuation, or uncertain. The verify field contains `subject_entity`, `transition_anchor`, and `grounding_entity`.
Verify entity format: Each unit contains `value`, `role`, `attributes`, `visibility`, and `importance`. Visibility is one of `visible`, `missing_in_final`, or `implied_from_caption`; importance is `required` or `soft`.
Output requirements: Output valid JSON only. `bridge_text` should be short, event-level, and mention or imply at least one subject and one required transition anchor.

Table 8: Prompt template for latent causal bridge planning and verify-unit extraction.

Rewrite and Verification Prompt
Instruction: Conservatively revise the previous ending. Keep as much of the original wording, names, objects, and concrete details as possible. Only make the smallest edit needed to fix a clear story-event mismatch.
Input: Previous ending `<initial_pred_response>`; latent bridge or bridge text; caption context; required content slots; emotion, causal, and ending prototypes.
Verification criteria: Required content slots must cover the central subject, trigger object or transition anchor, and trigger source or grounding entity when present. Narrative slots check causal-trigger and emotion/ending prototypes with fuzzy or semantic matching.
Rewrite constraints: Do not replace specific details with vague generic phrases. Do not add unsupported events. Preserve the original prediction unless the causal card identifies a missing event, missing participant, or wrong event direction.
Output requirements: Output only story text. Use one to three natural sentences depending on the dataset setting.
Diagnostics: Store verification result, missing slots, rewrite attempts, repair candidates, and whether the content-overlap guard falls back to the initial prediction.

Table 9: Prompt template for constraint-aware story rewriting and verification.