STORYRELAY: REFERENCE-ADDRESSABLE AUTOREGRESSIVE VISUAL STORYTELLING

Anonymous authors
Paper under double-blind review

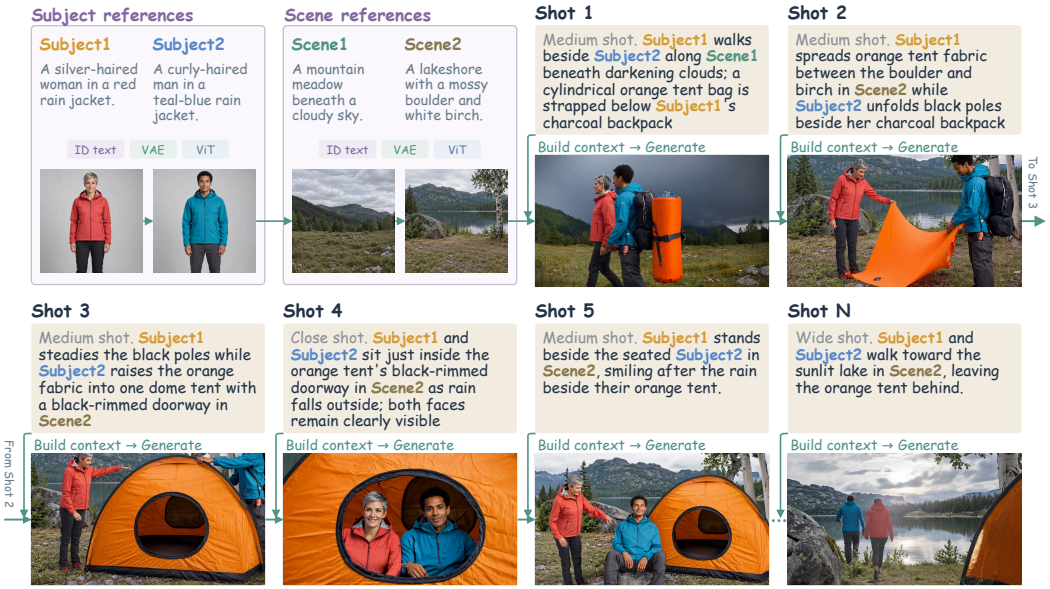


Figure 1: Reference-addressable autoregressive visual storytelling. Given persistent subject and scene references together with an ordered sequence of shot descriptions, StoryRelay generates the story one shot at a time. Each new shot conditions on the relevant original references and selected previous outputs, allowing subject identity and visual state to carry forward across the story.

ABSTRACT

Reference-conditioned visual storytelling must maintain subject identity and visual continuity across a sequence of shots while adapting to changing actions, viewpoints, and compositions. Generated history can improve cross-shot continuity, but it may also introduce irrelevant or incorrect visual state into later generations. We introduce **StoryRelay**, a reference-addressable autoregressive storytelling framework that jointly leverages persistent references and evolving visual history across shots. StoryRelay maintains story-level reference bindings and allows shot descriptions to address subject references and optional scene references through textual identifiers. For autoregressive story learning, *Autoregressive Story Modeling* (ASM) post-trains a unified multimodal model on interleaved reference–text–image stories, predicting each target keyframe from reference images and preceding ground-truth history. To strengthen identity resolution across multiple visual sources, *Shot-Aware Identity Grounding* (SIG) associates requested subjects with identity-specific evidence from references and historical appearances. At inference, a *feedback-controlled autoregressive rollout* performs history selection, generation, verification, and bounded shot-local correction within the same model. On ViStoryBench, StoryRelay improves cross-reference identity similarity from 0.45 to 0.52 and character-count matching from 65.57 to 68.46 over the original BAGEL.

1 INTRODUCTION

Visual storytelling turns a script into an evolving sequence of images. When reference images specify recurring subjects and, optionally, scenes, each new shot draws on two complementary visual contexts:

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



Figure 2: Original BAGEL fails to compose the requested characters correctly. Adding ASM, SIG, and feedback-controlled rollout progressively improves character identity preservation.

persistent references that specify recurring entities and preceding shots that record how the story has visually evolved. Generating shots independently may satisfy each reference and shot text, yet fail to preserve identity and visual continuity across the sequence. Conversely, using generated history can improve continuity, but may propagate irrelevant context or generation errors into later shots. Figures 2 and 3 illustrate identity substitutions and duplicate subjects.

Recent visual storytelling methods propagate information across shots through autoregressive image–text context, shared cross-frame features, or selected visual history (Liu et al., 2024; Yang et al., 2025a; Zhou et al., 2024; An et al., 2026). We consider a *reference-addressable autoregressive storytelling* setting, in which persistent story-level identifier-to-reference bindings and evolving visual history jointly condition each next shot. Unified multimodal models provide a natural foundation because they can consume and generate interleaved text and images (Deng et al., 2025). StoryRelay maintains a story-level reference registry from which each generation call retrieves the required assets. Shot descriptions address these assets through textual identifiers such as `Subject1` and `Scene1`.

StoryRelay realizes reference-addressable storytelling by keeping story-level references persistently accessible while relaying visual state from earlier to later shots. It implements this design through three complementary mechanisms: a representation and training interface, identity-grounding supervision, and inference-time state management. *Autoregressive Story Modeling* (ASM) trains next-keyframe prediction from interleaved references, text, and teacher-forced clean history. Because serialization alone does not resolve competing identities and visual sources, *Shot-Aware Identity Grounding* (SIG) distinguishes requested from unused identities through spatial supervision. At inference, a *feedback-controlled autoregressive rollout* selects relevant history, verifies generated candidates, and admits only accepted outputs into subsequent history. Together, these mechanisms maintain persistent reference access while controlling the visual history passed to later shots.

Training ASM and SIG requires cross-shot correspondence among textual identifiers, visual references, and recurring entities. We therefore construct identity-aligned video-to-story supervision, yielding 14,565 stories that link ordered keyframes, subject and scene references, reference-conditioned training annotations, and spatial masks. Training uses target-derived shot annotations as conditioning text, whereas evaluation uses the benchmark-provided prompts as supplied. On ViStoryBench, StoryRelay improves subject-reference identity and character-count matching over pretrained BAGEL. Overall, StoryRelay combines reference-addressable story modeling, identity-specific grounding across original and historical visual sources, and verification-controlled admission of generated history. Our contributions are:

- **Autoregressive Story Modeling (ASM).** We introduce a reference-addressable multimodal story representation that combines persistent story-level reference bindings with interleaved visual history.

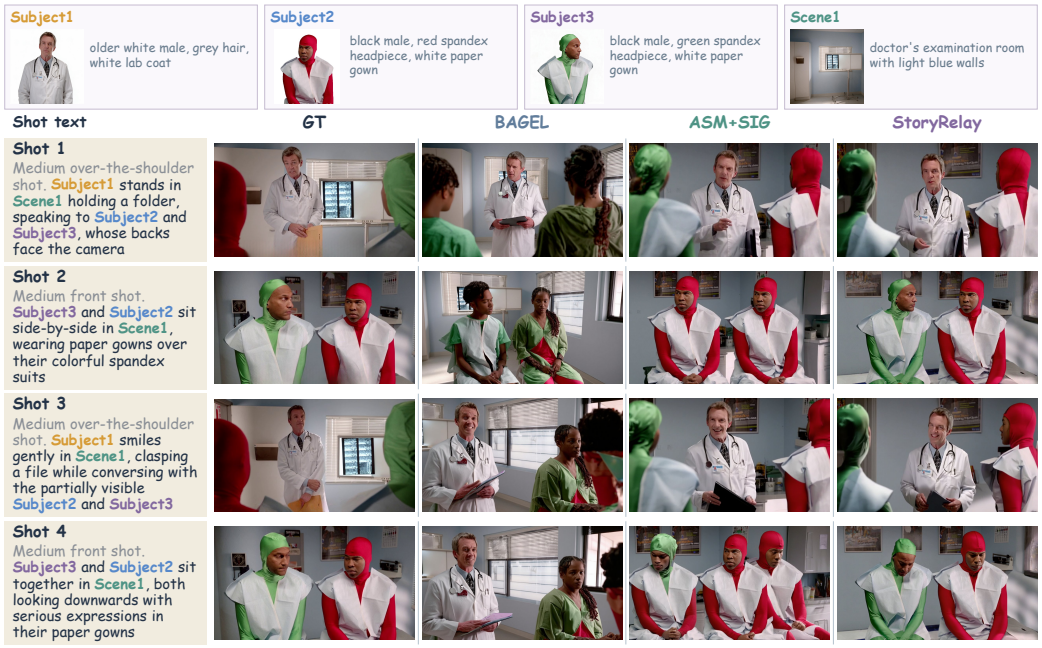


Figure 3: Original BAGEL copies visual content from Shot 2 into later shots, causing duplicated characters and identity drift. StoryRelay avoids such history copying while preserving character identities across shots.

- **Shot-Aware Identity Grounding (SIG).** We introduce identity-specific spatial supervision across original references and historical appearances, improving grounding of requested subjects when multiple identities and visual sources coexist in the autoregressive context.
- **Feedback-controlled autoregressive rollout.** We reuse the same multimodal model for history selection, generation, verification, and shot-local correction, while admitting only accepted outputs into the history available to subsequent shots.

2 Related Work

Story context and reference retrieval. Maintaining continuity across shots requires a model to exploit prior visual context while preserving access to persistent references. StoryGen and SEED-Story generate stories autoregressively from image–text context (Liu et al., 2024; Yang et al., 2025a), while OneStory selects historical frames for next-shot generation (An et al., 2026). StoryDiffusion, One-Prompt-One-Story, and Story-Iter promote cross-shot consistency through shared attention, coupled prompts, or iterative refinement (Zhou et al., 2024; Liu et al., 2025; Mao et al., 2026), and IdentityStory discovers and injects identities for multi-character stories (Zhou et al., 2026). VstoryGen combines character and background references with shot-type control, using a keyed visual memory bank to retrieve static references and recent generated frames with matching characters and scenes (Li et al., 2026). StoryRelay also combines original references and generated history, with ASM for autoregressive story training, SIG for cross-source identity supervision, and candidate-checking feedback to govern history admission.

Spatial grounding across visual sources. When multiple subjects and visual sources coexist, reliable generation depends on grounding each requested identity to the correct visual evidence. DreamO aligns image customization through reference response–mask supervision (Mou et al., 2025), while OmniGen2 supports reference-conditioned generation and editing (Wu et al., 2026). StoryRelay extends spatial grounding to autoregressive storytelling through SIG, which associates requested identities with regions from both original references and historical appearances. This provides identity-specific supervision when multiple subjects and visual sources coexist in the same story context.

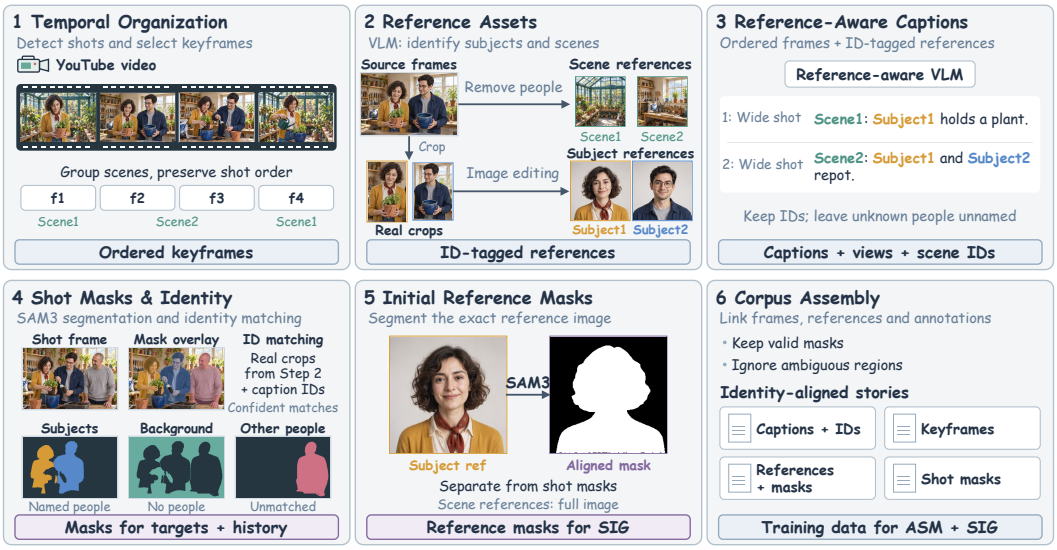


Figure 4: Six-stage identity-aligned story construction. Edited reference assets condition generation, while retained real crops support identity matching; ordered annotations, keyframes, and masks are assembled into supervision for ASM and SIG.

Feedback-controlled generation. Autoregressive generation can propagate errors through history, making feedback especially useful for deciding whether generated content should influence later shots. BAGEL provides a unified interface for interleaved multimodal understanding and generation (Deng et al., 2025). Prior feedback systems use model feedback to revise outputs or coordinate generation and editing tools (Yang et al., 2025b; Wang et al., 2024; Cheng et al., 2026). ToolArtist integrates native image generation and iterative feedback within a unified multimodal policy (Zhao et al., 2026). In autoregressive storytelling, StoryRelay uses candidate verification to control which generated shots enter the visual history available to subsequent shots.

3 Data Construction

We convert YouTube videos into identity-aligned stories through the six-stage pipeline in Figure 4. The resulting stories pair ordered keyframes, reference assets, target-derived training annotations, and spatial annotations to provide interleaved supervision for ASM and identity-linked regions for SIG.

3.1 Ordered Keyframes and Identity References

Temporal organization (Stage 1). We detect shot boundaries, select one sharp keyframe per shot, and group consecutive shots into bounded-length stories while preserving temporal order.

Reference assets (Stage 2). A video-level VLM identifies recurring subjects and scenes and assigns shared identifiers such as `Subject1` and `Scene1`. We retain real subject crops for identity matching and edit them into isolated references for generation; scene references are edited to remove people while preserving the environment.

3.2 Reference-Conditioned Captions and Spatial Annotations

Reference-conditioned training annotations (Stage 3). The annotation VLM jointly observes identifier-tagged references and ordered keyframes and produces per-shot descriptions, camera views, and scene identifiers with consistent subject IDs. These texts are target-derived training annotations; ViStoryBench evaluation instead uses the benchmark-provided prompts unchanged.

Shot masks and identity matching (Stage 4). Offline SAM 3 segmentation (Carion et al., 2026) proposes person masks, which are matched to named subjects using retained real crops and multimodal

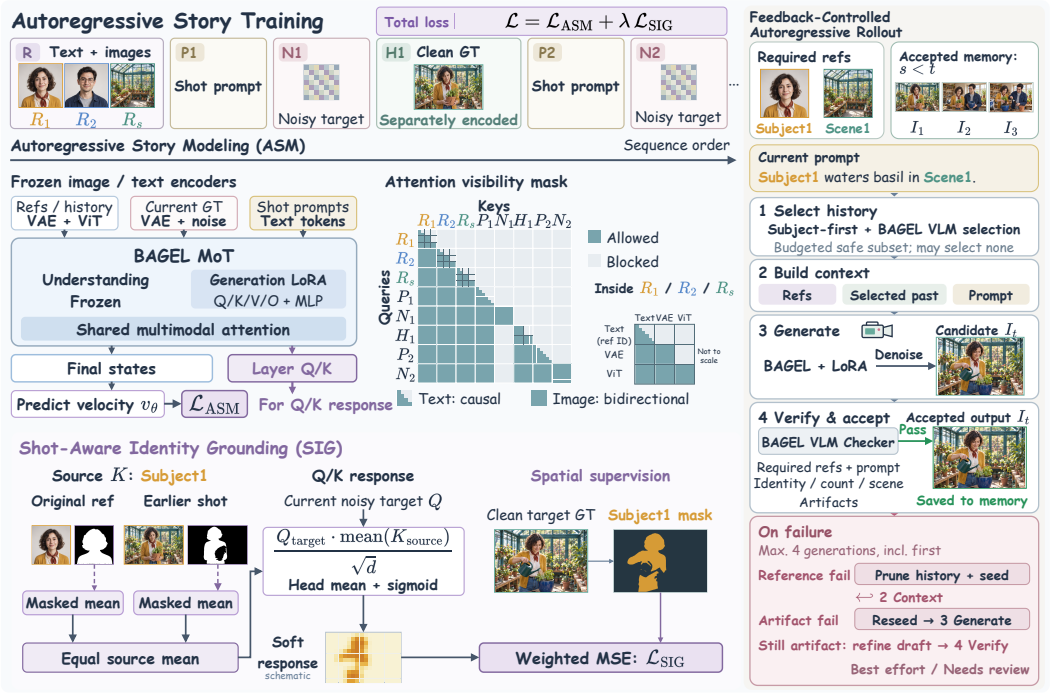


Figure 5: StoryRelay: ASM and SIG training (left); feedback-controlled autoregressive rollout (right).

similarity cues. Confident matches define subject regions; all detected people are excluded from scene/background targets. These regions supervise the current shot and provide identity-specific sources when the corresponding keyframes are used as history during training.

Initial-reference masks (Stage 5). We separately segment the exact subject-reference image used for conditioning rather than reusing a crop or shot mask. These masks define initial subject-source regions for SIG, while scene references use the full image. If a mask is invalid, we exclude only that auxiliary SIG source while retaining the reference for generation conditioning.

3.3 Corpus Assembly and Statistics

Aligned training stories (Stage 6). We assemble references, ordered keyframes, training annotations, and valid spatial regions into identity-aligned stories. ASM uses the interleaved context for next-keyframe modeling, while SIG uses identity-linked source and target regions for auxiliary supervision without requiring masks at inference. The corpus contains 14,565 stories and 53,418 keyframes from 8,907 videos, averaging 3.67 keyframes per story with a range of 2–12. Appendix A provides implementation details, annotation checks, and full corpus statistics.

4 Method

Given reference assets $\mathcal{R} = \{(e, r_e)\}$, comprising subject images and optional scene images, and shot texts $c_{1:T}$ with recurring entity identifiers, StoryRelay generates keyframes as $x_t \sim p_\theta(\cdot \mid c_t, \mathcal{R}_t, H_t)$, where $\mathcal{R}_t \subseteq \mathcal{R}$ contains required references and H_t is selected from previously accepted shot outputs. The three parts of this conditioning play distinct roles: $\mathcal{R}_t$ specifies what the recurring entities look like, H_t carries the visual state the story has reached, and c_t requests the next transition. Figure 5 summarizes training on the aligned stories of Section 3 and the feedback-controlled autoregressive rollout with BAGEL (Deng et al., 2025).

During training, ASM uses a shared reference prefix and teacher-forced clean history. At inference, each shot reassembles the required references and selected accepted history into a fresh conditioning context, while story-level reference bindings persist across the sequence.

Story-level bindings and per-call context. Story-global entity identifiers and their reference assets are maintained externally. For each generation call, the required references and selected history are assembled into a local context, and global identifiers are deterministically remapped to a consistent per-call local namespace. The remapping changes only the labels used within that call, while preserving the underlying entity–reference bindings.

Throughout, c_t denotes the shot text: a target-derived annotation during training and the benchmark-provided prompt at evaluation. Evaluation prompts are used unchanged, without target-conditioned recaptioning.

4.1 Autoregressive Story Modeling

ASM serializes each training story with one reference prefix, in which every local identifier, such as `Subject1`, is associated with its visual reference. The mapping is fixed within this training sequence, and successive shot texts reuse it to address the same entity. Training is teacher-forced: preceding ground-truth keyframes provide clean history, while noise is applied only to generation targets. ASM therefore learns next-keyframe prediction from identity-bound context.

Let z_t^τ denote the noisy target latent at flow time τ, and h_t the ground-truth keyframe's clean VAE latents and ViT features. Reference images provide the same two visual representations alongside their identity labels. A training story is serialized as

$$[\mathcal{R}]\ [c_1, z_1^\tau, h_1]\ [c_2, z_2^\tau, h_2]\ \cdots\ [c_T, z_T^\tau]. \tag{1}$$

In Equation 1, $[\mathcal{R}]$ denotes the story-wide training reference prefix: all enabled references used in that training story, subject to the dropout described in Appendix B. It is shared across target positions, not filtered separately for each shot. At inference, $\mathcal{R}_t \subseteq \mathcal{R}$ instead denotes the references required for the current shot. Here z_t^τ is the noisy generation target, whereas h_t is a separately encoded clean copy exposed only to later shots as history. Each target reads the reference prefix, preceding clean history, and current shot text; its own clean copy, future images, and earlier noisy targets are masked out. Tokens within an image interact bidirectionally, and packed stories remain isolated. We retain BAGEL's latent flow-matching objective $\mathcal{L}_{\mathrm{ASM}}$ (Lipman et al., 2023); packing and positional details are given in Appendix B.2. This motivates the inference-time state management in Section 4.3, which controls which generated outputs can condition later shots.

4.2 Shot-Aware Identity Grounding

Serializing reference identifiers alone does not ensure correct identity resolution when multiple identities and visual sources coexist in the same autoregressive context. SIG reinforces that grounding, extending DreamO's response–mask alignment (Mou et al., 2025) to shot-conditioned identity sources. Identity-specific masks separate co-occurring subjects; within-source normalization and equal-source pooling prevent larger visible regions from receiving more weight simply because they contain more tokens; shot-conditioned targets distinguish requested identities from unused ones. Spatial masks are training-only.

For entity e at shot t, the target $Y_{t,e}$ is the foreground mask for a mentioned subject or the background mask for a mentioned scene. Available identities omitted from the shot text receive zero-response targets; invalid annotations are skipped.

Identity sources and equal-source pooling. For identity e, $\mathcal{S}_{t,e}$ contains eligible original references and earlier clean-history sources. Let $m_{s,e}(v)$ be source-mask coverage at token v and $K_{s,v}^{\ell,h}$ a normalized, rotary-encoded key. We first pool within each source and then average sources:

$$\bar{K}_{s,e}^{\ell,h} = \frac{\sum_v m_{s,e}(v) K_{s,v}^{\ell,h}}{\sum_v m_{s,e}(v)}, \qquad \bar{K}_{t,e}^{\ell,h} = \frac{1}{|\mathcal{S}_{t,e}|} \sum_{s \in \mathcal{S}_{t,e}} \bar{K}_{s,e}^{\ell,h}. \tag{2}$$

Equal-source pooling gives each source equal total weight, independent of its visible area or token count. Appendix B.3 specifies source eligibility and target weighting.

Spatial response alignment. For target query $Q^{\ell,h}_{t,u}$ after the same normalization and rotary encoding, the identity response is

$$P^{\ell}_{t,e}(u) = \sigma\left(\frac{1}{H}\sum_{h=1}^{H}\frac{\langle Q^{\ell,h}_{t,u}, \bar{K}^{\ell,g(h)}_{t,e}\rangle}{\sqrt{d_h}}\right), \tag{3}$$

where σ is sigmoid, H the query-head count, d_h the head dimension, and $g(h)$ the grouped-key head mapping. This pooled response is used only as an auxiliary training signal; it does not replace native attention or impose a separate routing constraint at inference.

Joint objective. The squared-error loss uses spatial/group weights $w_{t,e,u}$ and supervised layers Λ:

$$\mathcal{L}_{\mathrm{SIG}} = \frac{\sum_{t,e,u} w_{t,e,u}\,|\Lambda|^{-1}\sum_{\ell\in\Lambda}(P^{\ell}_{t,e}(u) - Y_{t,e}(u))^2}{\max(1, \sum_{t,e,u} w_{t,e,u})}. \tag{4}$$

The combined objective $\mathcal{L} = \mathcal{L}_{\mathrm{ASM}} + \lambda\mathcal{L}_{\mathrm{SIG}}$ optimizes only generation-branch attention and MLP LoRA parameters (Hu et al., 2022); the base model remains frozen. We use $\lambda = 0.1$ by default. Appendix B gives the supervised layers, weights, and optimization settings.

4.3 Feedback-Controlled Autoregressive Rollout

At inference the history H_t contains generated rather than ground-truth frames, so unreliable outputs can carry errors into later shots. The rollout therefore manages the autoregressive state itself: it selects history, verifies candidates, and applies bounded shot-local correction, admitting only outputs that pass verification to subsequent history.

History selection. The rollout reuses BAGEL for history selection and verification, with rule-based retry decisions rather than a separately trained controller. For each shot, it first filters accepted history by requested-subject overlap and then selects a relevant subset, possibly empty, under the context budget. Required references are always retained, and selected frames are restored to temporal order. For attempt a of shot t, the subset $H_{t,a}$ and its local-ID map $\pi_{t,a}$ define

$$C_{t,a} = \mathrm{Serialize}_{\pi_{t,a}}(\mathcal{R}_t, H_{t,a}, c_t). \tag{5}$$

The multimodal context is rebuilt after each history update rather than appended to a single story-wide model context.

Verification and correction. BAGEL checks each candidate against required references and the shot text, focusing on reference consistency and severe visual defects rather than full narrative correctness. Reference failures (identity, count, binding, or scene mismatch) trigger history pruning and reseeding; artifact failures trigger reseeding followed, when appropriate, by draft refinement. Correction is shot-local: completed shots are not revised. The budget is at most four generation attempts including the initial candidate, with early stopping on a pass or branch termination.

History admission. History stores only accepted shot outputs: the first candidate that passes verification is accepted and appended. If no candidate passes within the attempt budget, BAGEL selects one best-effort output solely to provide an image for the current shot. This fallback retains its failure/review flags, is not added to history, and cannot condition subsequent shots. All final outputs, including fallback images, are evaluated. Appendix B specifies decisions, identifier remapping, token budgets, retry branches, and parsing fallbacks (Table 7); Appendix D traces a rollout.

5 Experiments

We evaluate subject-reference fidelity, within-story identity and style consistency, character-count matching, image quality, and prompt alignment on ViStoryBench (Zhuang et al., 2026); Appendix C records the protocol and provenance.

Table 1: ViStoryBench comparison. Published results appear above our local runs; bold marks the better of our two runs.

Method / run	CIDS-cross ↑	CIDS-self ↑	CSD-cross ↑	CSD-self ↑	OCCM ↑	IS ↑	Aes. ↑
StoryGen	0.4140	0.5930	0.2830	0.5800	41.10	7.31	3.740
StoryDiffusion	0.4360	0.5650	0.3400	0.5470	57.40	10.06	5.130
Story-Adapter $s = 0$	0.4600	0.6050	0.4560	0.5480	69.00	12.98	4.990
Story-Adapter $s = 5$	0.4010	0.6260	0.3250	0.7370	59.30	13.73	4.890
UNO (Less-to-More)	0.4850	0.6200	0.3910	0.6020	74.20	12.40	5.230
OmniGen2	0.5480	0.6470	0.4540	0.6000	70.20	11.05	5.250
CharaConsist	0.3150	0.5190	0.2820	0.5530	57.90	13.80	5.880
QwenImageEdit-2509	0.4750	0.5740	0.3810	0.5930	59.80	13.42	5.500
BAGEL	0.4549	0.5816	0.4232	0.5604	65.57	10.59	**4.639**
StoryRelay	**0.5200**	**0.5876**	**0.4450**	**0.5735**	**68.46**	**12.51**	4.632

5.1 Experimental Setup

Benchmark and metrics. We use all 80 English stories and 1,317 target shots with the official character-type tags, multi-face identity encoders, and metric implementations. Metrics cover subject-reference and within-story identity (CIDS), style similarity (CSD), onstage character count matching (OCCM), IS, aesthetic score (Aes.), Copy-Paste (CP), and prompt alignment (PA). PA averages scene, camera, global-action, and local-action scores on a 0–100 scale. The benchmark uses character reference images and textual scene descriptions; scene PA measures alignment with the described setting. Appendix D illustrates conditioning on visual scene references.

Models. BAGEL is pretrained without an adapter; ASM is story-trained without SIG; ASM+SIG adds identity grounding; StoryRelay combines ASM+SIG with feedback-controlled autoregressive rollout. Appendix C records configuration differences; Appendix B gives the training configuration (Table 5).

Generation protocol. Our runs share story inputs, reference assets, 1024×576 resolution, 50 time points / 49 integration steps, and a four-image history cap. Single-pass runs use recent generated history; StoryRelay selects accepted outputs under Section 4.3's policy. All final outputs are scored. No ViStoryBench story was used for training or controller development; Appendix C records the data splits and cross-source protocol details.

5.2 ViStoryBench Comparison

Table 1 compares identity, style, character count, and image quality with the published results from Zhuang et al. (2026). Appendix C details the evaluation protocols and result provenance.

Reference fidelity and count. On the same BAGEL backbone, StoryRelay improves CIDS-cross from 0.4549 to 0.5200 ($+14.31\%$ relative) and OCCM from 65.57 to 68.46 ($+2.89$ points), showing stronger reference identity preservation and character-count matching.

Prompt alignment and other metrics. Table 2 reports the four PA components for locally evaluated BAGEL and StoryRelay. Mean PA increases from 49.44 to 60.62, while within-story identity metrics change only modestly and the aesthetic score remains close to pretrained BAGEL. The rollout verifier checks reference consistency and severe visual defects rather than optimizing the PA criteria directly.

5.3 Configuration Analysis

Table 3 compares four configurations on the BAGEL backbone under the same local evaluation protocol.

Table 2: Prompt alignment for BAGEL and StoryRelay on ViStoryBench (0–100; higher is better).

Method	Scene	Camera	Global act.	Local act.	PA
BAGEL	42.04	57.53	56.32	41.88	49.44
StoryRelay	**63.99**	**57.98**	**67.40**	**53.12**	**60.62**

Table 3: Local configuration comparison on ViStoryBench. Bold marks the best configuration for each metric.

Method	CIDS ↑		CSD ↑		OCCM ↑	IS ↑	Aes. ↑	CP ↓	PA ↑
	Cross	Self	Cross	Self					
BAGEL	0.4549	0.5816	0.4232	0.5604	65.57	10.59	**4.639**	0.5078	49.44
ASM	0.4573	0.5836	0.4269	0.5729	65.78	10.64	4.613	0.5009	52.43
ASM+SIG	0.4735	0.5848	0.4355	**0.5740**	66.22	10.97	4.528	0.5035	52.60
StoryRelay	**0.5200**	**0.5876**	**0.4450**	0.5735	**68.46**	**12.51**	4.632	**0.4984**	**60.62**

Method	Scene ↑	Camera ↑	Global act. ↑	Local act. ↑	PA ↑
BAGEL	42.04	57.53	56.32	41.88	49.44
ASM	41.59	57.51	63.41	47.19	52.43
ASM+SIG	42.38	57.63	62.92	47.50	52.60
StoryRelay	**63.99**	**57.98**	**67.40**	**53.12**	**60.62**

Sequential configuration comparison. Across the sequential configurations, CIDS-cross increases from 0.4549 for BAGEL to 0.4573 for ASM, 0.4735 for ASM+SIG, and 0.5200 for StoryRelay. ASM+SIG and StoryRelay share the same adapter checkpoint and differ at inference in the rollout policy. CSD-self is essentially unchanged between these two settings (0.5740 vs. 0.5735). Aesthetic score drops to 4.528 under ASM+SIG but returns to 4.632 under StoryRelay, close to the pretrained BAGEL value of 4.639. The PA increase is concentrated most strongly in scene alignment, which rises from 42.38 to 63.99 between ASM+SIG and StoryRelay, with smaller changes in global and local actions and little change in camera alignment.

Feedback-controlled autoregressive rollout. Scene information does enter the rollout: history selection scores scene overlap alongside subject overlap, and verification treats scene mismatch as a reference failure alongside identity and count. The observed scene-PA increase is consistent with this behavior, though the aggregate PA results do not isolate individual controller branches. Appendix D.3 traces history selection and correction.

5.4 Qualitative Evidence

Figures 2–3 and the cases in Appendix D are curated illustrations from independent rollouts; all quantitative performance conclusions are based on ViStoryBench.

The early comparisons separate two failure modes. In Figure 2, Shot 3 has the correct count under ASM+SIG but substitutes a previous character for the requested doctor. In Figure 3, Shots 2 and 4 duplicate a patient or add an extra person. StoryRelay retains the requested pair in these views. Appendix D provides shot-level analysis, four further stories, and a controller trace.

6 Conclusion

StoryRelay combines persistent reference grounding with feedback-controlled history management for autoregressive visual storytelling. Original references provide stable anchors for recurring entities, while generated history carries evolving visual state. On ViStoryBench, StoryRelay improves subject-reference identity similarity, character-count matching, and prompt alignment over the pretrained BAGEL backbone, with the largest alignment gain in scene correspondence. These results support an approach that preserves access to original references while selectively admitting and using generated history. The central design principle is to treat generated history as state to be managed, rather than context to be trusted unconditionally.

AI USE STATEMENT

We used generative AI tools to assist with literature organization, code-to-method cross-checking and structural editing, and LaTeX formatting. Generative AI models were also used in training-data construction, including reference-image editing and the generation of reference-conditioned shot annotations for identity-aligned video-to-story training examples. The resulting data were further processed by the data-construction pipeline described in Section 3. AI-assisted checks do not replace author verification. The authors take responsibility for the final manuscript and research results, including technical claims, derivations, experimental decisions, and AI-assisted training-data construction.

ETHICS STATEMENT

StoryRelay uses pretrained generative models and supervision derived from YouTube videos. Reference-conditioned synthesis may inherit biases from the underlying models, source videos, and automatic annotations, and can enable misleading depictions of identifiable people. Automated verification and improvements in consistency scores do not by themselves ensure correct, safe, or unbiased outputs. Responsible use calls for attention to source licenses, privacy, and consent, together with provenance tracking, content-safety measures, and human review of generated outputs.

REPRODUCIBILITY STATEMENT

Data construction, annotation checks, and corpus statistics are described in Appendix A. Training objectives, trainable modules, hyperparameters, history selection, and inference budgets are specified in Appendix B. Evaluation protocols, evaluator configurations, and result provenance are reported in Appendix C.

REFERENCES

Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, Chenyang Zhang, Tao Xiang, Fanny Yang, Serge Belongie, and Tian Xie. Onestory: Coherent multi-shot video generation with adaptive memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16173–16184, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/An_OneStory_Coherent_Multi-Shot_Video_Generation_with_Adaptive_Memory_CVPR_2026_paper.html.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Junhao Cheng, Xi Lu, Hanhui Li, Khun Loun Zai, Baiqiao Yin, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. AutoStudio: Crafting Consistent Subjects in Interactive Story Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4770–4779, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/AISTORY/html/Cheng_AutoStudio_Crafting_Consistent_Subjects_in_Interactive_Story_Generation_CVPRW_2026_paper.html.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging Properties in Unified Multimodal Pretraining. *arXiv preprint arXiv:2505.14683*, 2025. URL https://arxiv.org/abs/2505.14683v3. Version 3; preprint.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Wei-Hua Li, Cheng Sun, and Chu-Song Chen. Customized Visual Storytelling with Unified Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4790–4800, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/AISTORY/html/Li_Customized_Visual_Storytelling_with_Unified_Multimodal_LLMs_CVPRW_2026_paper.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent Grimm - Open-ended Visual Storytelling via Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6190–6200, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Intelligent_Grimm_-_Open-ended_Visual_Storytelling_via_Latent_Diffusion_Models_CVPR_2024_paper.html.

Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-Prompt-One-Story: Free-Lunch Consistent Text-to-Image Generation Using a Single Prompt. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cD1kl2QKv1.

Jiawei Mao, Xiaoke Huang, Yunfei Xie, Yuanqi Chang, Mude Hui, Bingjie Xu, Zeyu Zheng, Zirui Wang, Cihang Xie, and Yuyin Zhou. Story-Iter: A Training-free Iterative Paradigm for Long Story Visualization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=puBVb9vTah.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, Mengtian Li, Mingcong Liu, Yunsheng Jiang, Shaojin Wu, Songtao Zhao, Jian Zhang, Qian He, and Xinglong Wu. DreamO: A Unified Framework for Image Customization. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12. ACM, 2025. doi: 10.1145/3757377.3763956. URL https://doi.org/10.1145/3757377.3763956.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. GenArtist: Multimodal LLM as an Agent for Unified Image Generation and Editing. In *Advances in Neural Information Processing Systems*, volume 37, pp. 128374–128395. Curran Associates, Inc., 2024. doi: 10.52202/079017-4077. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e7c786024ca718f2487712bfe9f51030-Abstract-Conference.html.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ziyi Xia, Ze Liu, Chaofan Li, Haoge Deng, Kun Luo, Bo Zhang, Jiajun Zhang, Dong Liu, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Towards Instruction-Aligned Multimodal Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21964–21975, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_OmniGen2_Towards_Instruction-Aligned_Multimodal_Generation_CVPR_2026_paper.html.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. SEED-Story: Multimodal Long Story Generation with Large Language Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1871–1881, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025W/HiGen/html/Yang_SEED-Story_Multimodal_Long_Story_Generation_with_Large_Language_Model_ICCVW_2025_paper.html.

Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2Img: Iterative Self-refinement with GPT-4V for Automatic Image Design and Generation. In *Computer Vision – ECCV 2024*, volume 15096 of *Lecture Notes in Computer Science*, pp. 167–184. Springer Nature Switzerland, 2025b. doi: 10.1007/978-3-031-72920-1_10. URL https://link.springer.com/chapter/10.1007/978-3-031-72920-1_10.

Jiahao Zhao, Xiaomin Yu, Zhongxiang Sun, Fengwei Teng, Chengwei Qin, Xiaobin Hu, Jun Xu, and Shuicheng Yan. ToolArtist: Tool-Using Unified Multimodal Models for Agentic Image Generation. *arXiv preprint arXiv:2608.04436*, 2026. URL https://arxiv.org/abs/2608.04436v1. Version 1; preprint.

Donghao Zhou, Jingyu Lin, Guibao Shen, Quande Liu, Jialin Gao, Lihao Liu, Lan Du, Cunjian Chen, Chi-Wing Fu, Xiaowei Hu, and Pheng-Ann Heng. IdentityStory: Taming Your Identity-Preserving Generator for Human-Centric Story Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13593–13601, 2026. doi: 10.1609/aaai.v40i16.38365. URL https://ojs.aaai.org/index.php/AAAI/article/view/38365.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110315–110340. Curran Associates, Inc., 2024. doi: 10.52202/079017-3501. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c7138635035501eb71b0adf6ddc319d6-Abstract-Conference.html.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, Xuanyang Zhang, Xianfang Zeng, Zhewei Huang, Gang Yu, and Chi Zhang. ViStoryBench: Comprehensive Benchmark Suite for Story Visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9455–9467, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhuang_ViStoryBench_Comprehensive_Benchmark_Suite_for_Story_Visualization_CVPR_2026_paper.html.

A Data Construction and Corpus Statistics

A.1 Pipeline components

The implementation separates video organization, reference creation, shot-text annotation, and mask preparation. Shot detection uses PySceneDetect's content-change detector with a histogram-difference fallback. Candidate frames are sampled within each shot using configurable boundary margins and ranked by a Laplacian-variance sharpness measure. Histogram-based scene grouping precedes the division of consecutive shots into stories with configurable length bounds.

Gemini 3.1 Pro identifies recurring subjects and scenes from sparse video-level observations and assigns identifiers within each source video. For each subject, the pipeline retains a source frame, crop coordinates, and a real crop; Seedream 5.0 edits that crop into an isolated, white-background reference. The same editing model produces scene references by removing people while retaining the environment. Reference-conditioned annotation jointly presents these references and ordered keyframes to Gemini 3.1 Pro, producing target-derived training shot texts. The annotation prompt requires consistent subject identifiers across shots and instructs the VLM not to assign a known identifier to an unmatched person.

A.2 Annotation checks and mask matching

Training-annotation responses are skipped if parsing fails or the returned shot count differs from the selected keyframe count. Shot entries are sorted by returned indices before text–keyframe pairing.

Shot-mask construction first proposes people and then matches them to named subjects using real reference crops. Pair scores combine cosine similarity of mask-pooled SAM 3 visual features, HSV color similarity, subject-description grounding, candidate salience, and segmentation confidence. Description grounding uses the overlap with description-prompted masks weighted by their confidence scores. One-to-one assignment maximizes the total matching score subject to a minimum pair-score threshold. A match is valid only when forbidding that pair reduces the best total assignment score by at least the assignment-margin threshold; otherwise its subject mask is marked invalid.

All detected people contribute to the non-background region, including people not confidently matched to named subjects. Pixels covered by multiple candidate masks are ignored during supervision. Thus *background*, *unnamed person*, and *ignored overlap* have distinct meanings.

Initial-reference masks are a separate product. The loader prioritizes available reference, source, and crop files; the mask record names the exact selected image. A path mismatch invalidates that auxiliary source. An invalid initial mask does not remove the normal reference image from conditioning, and an entity may obtain valid historical sources when it reappears in subsequent shots. Mask-coverage and training-time validity checks are specified in Appendix B.

A.3 Corpus statistics

Table 4 reports the September 15, 2026 training snapshot. The mean length is 3.67 keyframes, computed as 53,418/14,565, and the longest story has 12 keyframes. Mask-validity counts are read from the stored annotation metadata.

Source videos are counted by a path-independent identifier, preferring a source UUID and falling back to the video filename stem. Story identifiers, training shot texts, frame indices, reference paths, and annotation records link each story to its source video.

B Implementation and Exact Inference Contract

B.1 Training configuration

Generation LoRA is attached to the generation-specific query, key, value, output, and gated-MLP projections. Base parameters remain frozen. The default SIG coefficient is $\lambda = 0.1$, also used for the evaluated SIG models; the architecture diagram denotes this weight symbolically as λ.

Table 4: Training corpus statistics. References are counted per story; image–subject pairs are deduplicated.

Quantity	Value
Stories / unique story identifiers	14,565 / 14,565
Unique source videos	8,907
Target keyframes	53,418
Shots per story (range)	2–12
Stories with 2 / 3 / 4 / 5 shots	950 / 8,356 / 2,610 / 1,263
Stories with 6 / 7 / 8 / 9 shots	649 / 344 / 228 / 103
Stories with 10 / 11 / 12 shots	46 / 10 / 6
Subject-reference occurrences	34,352
Occurrences marked mask-valid / invalid	34,127 / 225
Unique image–subject pairs	21,068
Unique pairs marked mask-valid / invalid	20,954 / 114

Table 5: Evaluated ASM+SIG training configuration.

Component	Setting
Backbone	BAGEL-7B-MoT
Trainable modules	Generation attention Q/K/V/O and MLP LoRA only
Trainable parameter count	161.5M
Adapter rank / scale / dropout	64 / 128 / 0
Frozen components	Base model, understanding path, ViT, VAE
Optimizer	AdamW, learning rate 5×10^{-5}, betas $(0.9, 0.95)$, weight decay 0
Schedule	Constant learning rate, zero warmup
Training implementation	Distributed data parallel, BF16, activation checkpointing, FlexAttention
Packing	Expected 80,000 tokens; maximum 108,920 per packed input
SIG supervision	Layers 7, 14, 21 (zero-based); sigmoid response; weighted MSE
Subject / scene coefficients	1 / 1
Unused-group mass coefficient	0.2; distinct from the overall SIG coefficient
Overall SIG coefficient	0.1 (default and evaluated setting)
CFG dropout	Text 0.1; VAE/ViT share an image draw at threshold 0.1
Reference descriptions	Omitted from generation context; identity labels retained
UNO / EMA	Disabled

Input format. Each image reference entry contains a textual identity label, such as `Subject1`, followed by its VAE and ViT image representations; the label is not an additional learned identity token. Natural-language appearance descriptions and the video-wide description are omitted from the training generation context. Local labels follow first occurrence in the training story's shot texts. Each historical block contains its shot text and clean image representations. Training reference and modality dropout are applied in the following order.

1. Reference-prefix sampling. One draw per story determines the initial reference prefix:

$$\text{initial reference prefix} = \begin{cases} \varnothing, & \text{with probability } 0.2, \\ \text{valid images with local-ID labels}, & \text{with probability } 0.8. \end{cases}$$

The empty branch omits all initial subject/scene reference images and their prefix labels, rather than replacing them with descriptions. With appearance descriptions disabled, the per-reference text/image-drop branch is inactive. These probabilities precede modality CFG and are not independent drops for each asset.

Table 6: SIG supervision targets.

Condition	Auxiliary supervision
Mentioned subject, valid source and mask	Foreground target for that subject.
Mentioned scene, valid source and mask	Background target excluding detected people.
Available source, entity not in shot text	Zero-response target with bounded unused-group weight.
Mentioned entity, untrustworthy target mask	Skip the target; do not interpret annotation failure as absence.
No valid causal source	Skip that routing group; retain ordinary generation supervision.

2. Modality CFG dropout. CFG sampling is separate from the reference-prefix draw and applies to CFG-enabled text and visual conditioning blocks, including retained reference images and clean history. Text is dropped with probability 0.1; when text is dropped, image dropout is disabled. Otherwise, a shared image draw at threshold 0.1 drops the VAE and ViT conditioning representations together. The effective image-CFG drop probability is therefore $0.9 \times 0.1 = 0.09$, not two independent 0.1 events. For valid enabled initial references, combining both stages gives image omission probability $0.2 + 0.8 \times 0.09 = 0.272$ and retention probability 0.728. These are derived sampling probabilities under the reported training configuration, not measured frequencies. Evaluation applies no stochastic reference dropout; the text/image-dropped guidance contexts serve classifier-free guidance instead.

Image transforms. The target/history VAE transform uses maximum/minimum image sizes 1024/512 with stride 16; its ViT counterpart uses 518/224 with stride 14. Initial references use VAE limits 512/512 and ViT limits 224/224, with aspect-ratio-aware resizing. Mask coverage is resized to the latent grid using area interpolation and aligned to token order. A 1024×576 target has a 64×36 latent grid, or 2,304 generation tokens, before references, understanding tokens, text, and special tokens are included.

B.2 ASM OBJECTIVE AND PACKING

For ground-truth keyframe $x_t^\star$, let z_t^0 be its clean VAE latent. With noise $\epsilon_t \sim \mathcal{N}(0, I)$ and flow time $\tau \in [0, 1]$, the noisy target is $z_t^\tau = (1 - \tau)z_t^0 + \tau\epsilon_t$. Conditioned on the visible reference, text, and clean-history context C_t, BAGEL's latent flow-matching objective is

$$\mathcal{L}_{\text{ASM}} = \mathbb{E} \left\| v_\theta(z_t^\tau, \tau, C_t) - (\epsilon_t - z_t^0) \right\|_2^2 . \tag{6}$$

The expectation is over training targets, noise, and flow time. In the serialization of Equation 1, noisy target blocks use the current sequence-level RoPE position without advancing its counter. This avoids extra positional offsets in subsequent context from training-only targets. One packed forward pass supervises multiple targets, while each clean historical block remains visible only to later targets.

B.3 AUXILIARY-LOSS DETAILS

A source is eligible only when its VAE tokens are present in the target's visible context and its required mask is valid and aligned with the loaded image. A dropped VAE source, missing source mask, or absent packed token range is excluded from the auxiliary source set defined in Section 4.2. A target with no eligible sources contributes no corresponding routing group. Subject sources use foreground masks matching the exact loaded image. Initial scene references use their entire latent grid; historical scenes use the background outside all detected people.

Uniform pooling over all masked tokens would weight sources in proportion to visible area. Equation 2 instead gives each source equal total weight, but does not reserve a fixed share for the original reference: with one reference and n historical views, history receives $n/(n+1)$ of the pooling weight. Original references remain explicit conditioning inputs. In Equation 3, query and key features are taken after normalization and rotary encoding; $g(h)$ maps query heads to their grouped key/value heads. Target queries read preceding source keys in causal order, without replacing ordinary attention.

For n_+ positive groups and n_- unused groups, each unused group receives weight $0.2 \max(n_+, 1)/n_-$ when $n_- > 0$ (Table 6).

For informative positive masks, target coverage below 0.005 is filtered; scene coverage above 0.98 is filtered as nearly uniform. Ignored overlap regions contribute weight $1 - I(u)$ at token position u. The negative-group scale is assigned during data construction, so later source dropout and filtering can alter realized positive/negative contributions. The SIG denominator is aggregated across distributed ranks; local numerators are scaled to compensate for distributed data-parallel gradient averaging. Ranks with no valid routing groups still participate in the collective. These details matter when reproducing the weighted objective in Equation 4.

B.4 History selection and budgets

The implementation supports up to eight coarse-filtered candidates; the evaluated configuration caps selected history at four images. Only earlier outputs that passed checking are eligible as history; failed best-effort outputs are excluded. The rule score gives subject overlap weight 6.0, scene overlap 0.5, an exact-subject-set bonus of 3.0, penalties of 2.0 for extra subjects and 0.5 for extra scenes, and a small recency tie-breaker. A same-scene frame with unrelated named subjects and no required-subject overlap is rejected. An unparseable history selection falls back to the rule-filtered order; an unparseable checker response is treated as rejection.

The estimated token ceiling is 108,920, with a 10% safety margin. Mandatory reference and target costs are checked first; an over-budget mandatory context raises an error. Optional history uses the remaining budget. Full, text-dropped, and image-dropped guidance contexts use separate caches.

For selected history H_t, estimated visual and text token costs satisfy

$$\widehat{C}(\mathcal{R}_t) + \widehat{C}_{\text{target}}(c_t) + \sum_{(c_s, x_s) \in H_t} \widehat{C}(c_s, x_s) \leq (1 - \eta)B, \tag{7}$$

where $B = 108{,}920$, $\eta = 0.1$, and $\widehat{C}_{\text{target}}(c_t)$ includes the current shot text and target image tokens. The planner packs frames greedily in selection order, then restores chronology.

For each attempt, first occurrence in chronological history texts followed by the current shot text determines consecutive local subject and scene IDs. The resulting map is applied consistently to reference labels and shot texts. The same story-global entity may receive different local labels across calls, without changing its reference binding. Required references and selected historical images are re-encoded as clean VAE latents and ViT features, and generation/guidance caches are initialized for that context. History pruning rebuilds both the local namespace and the caches, removing the pruned frames' features. Persistent reference bindings do not imply a story-wide KV cache.

Only required subject reference images enter the generation prefix. Other retained history images may depict additional entities and participate in local namespace construction. Scene conditions use supplied images when available; the ViStoryBench evaluation supplies scene descriptions as text.

B.5 Generation, retries, and logging

Generation uses text/image guidance scales 4.0/1.5, guidance interval $[0.4, 1.0]$, timestep shift 3.0, and 50 time points / 49 integration steps. The output size is 1024×576.

The checker returns a pass flag, a failure class (none, reference, or artifact), failed reference identifiers, and a short repair instruction, rather than a calibrated quality score. Reference failures include missing/extra subjects, identity or binding errors, and incompatible scenes; severe visual defects are artifact failures.

Reference retries remove one history frame at a time by default. The maximum text budget is 512 tokens for checker and selection calls, split between reasoning and structured output. History-selection thumbnails use size 224. The four-attempt ceiling includes the initial generation; the artifact branch can stop after three candidates.

When no candidate passes, the all-failed selection receives every generated attempt from that shot, the current shot text, the required references, and the prior failure labels. BAGEL returns a structured `chosen_attempt` identifier, prioritizing subject identity/count/binding and required scene before

Table 7: Feedback-controlled autoregressive rollout actions.

Event	Action
First checker pass	Stop, return candidate, and append it to history with a passing verdict.
Reference failure	Remove low-priority/confusing history and change seed; use no history on the final reference retry. Mandatory original references remain.
Artifact failure	Retry the same context with another seed; if still appropriate, refine a draft with a short repair instruction. Prune optional history if the draft exceeds budget.
All attempts fail	BAGEL VLM selects the best available candidate as the current shot's output, retaining its failure/review flags. It is not appended to history; the story continues using previously accepted history.
Optional best-of-N	Disabled by default. If enabled, expand after the first pass with fixed context and select only among passing candidates.

visual quality. Selection ranks only the available candidates and does not turn a failed checker verdict into a pass. With one available candidate, it returns that candidate without a VLM comparison; if the comparison cannot be parsed or selects an invalid attempt, the implementation returns the earliest candidate. In all cases the result remains `accepted=false` and `needs_review=true`. The selected image is returned only to ensure an output for the current shot; neither it nor the other failed attempts enter memory or condition later shots. Selection metadata records the eligible attempts, chosen attempt, and VLM response or parsing error. Unlike optional best-of-N, this fallback does not generate extra candidates.

Historical context selection runs at the start of a shot; retries adjust the resulting plan rather than rerunning the selector. An initial reference failure enters history-pruning retries, with no history on the final reference retry. An artifact failure first triggers reseeding; draft refinement follows only if the updated decision still reports an artifact. Checker-identified reference failures are mapped from local identifiers back to global identities for history pruning.

Every attempt retains its image, context plan, checker response, and selected-history information. The final result records checker acceptance and review status separately, and all final images are retained for evaluation. An accepted shot output is one that passed the checker; selecting a failed fallback as the current shot's output does not make it accepted or eligible for history.

C Evaluation Protocol and Provenance

C.1 Benchmark inputs and metric implementations

The CIDS, CSD, OCCM, and IS results in Table 3 use all 80 English ViStoryBench stories and their 1,317 target shots. Original character references and benchmark-provided shot prompts are retained; scene conditions are textual descriptions, not reference images. All reported runs share the same scoring-code, environment, and metric-weight fingerprints, including the official character-type tags and multi-face identity encoders. CIDS measures cross-reference identity similarity and within-story identity consistency; CSD measures style similarity and consistency using a style-trained CLIP encoder; OCCM measures onstage character counts; IS uses the benchmark implementation. The local-configuration results are reported in Table 3; Table 2 provides the corresponding PA breakdown. Table 1 repeats the BAGEL and StoryRelay results only for cross-method context.

Original BAGEL uses the pretrained weights without an adapter; ASM and ASM+SIG are story-trained configurations, and StoryRelay adds feedback-controlled autoregressive rollout to ASM+SIG. Table 1 places our results alongside the published methods, and Table 3 in Section 5.3 compares the four configurations. The project README records exact run directories and checkpoint mappings.

C.2 Generation and aggregation

The output resolution is 1024×576, and history is capped at four images. The single-pass policy uses recent generated history; the evaluated feedback controller selects from accepted history and uses

failure-conditioned retries. Only the first passing candidate is appended to history. If all attempts fail, a best-effort image is returned for the current shot with its failure/review flags intact, but is excluded from subsequent history. Its attempt cap is four, including the initial candidate, with additional best-of-N disabled. Original reference images, inputs, output coverage, and scoring summaries have recorded fingerprints. Final outputs are scored in full: no image is dropped from the evaluated set on the basis of the rollout checker's flags. ASM+SIG and StoryRelay share an adapter checkpoint.

Reported identity, style, and character-count metrics average the native per-story aggregates over all 80 stories; IS uses the benchmark evaluator output and is not reconstructed from subsets. For the StoryRelay configuration versus original BAGEL, the CIDS-cross difference is $+0.0651$ and the OCCM difference is $+2.89$ points at the precision displayed in the tables.

C.3 Development discipline and data splits

No ViStoryBench story was used for model training, for developing the controller, for threshold selection, or for prompt design; the benchmark was used only for the final evaluation reported here. The controller's rule weights are fixed constants: subject overlap 6.0, scene overlap 0.5, an exact-subject-set bonus of 3.0, penalties of 2.0 for extra subjects and 0.5 for extra scenes, and a small recency tie-breaker. They were set before any evaluation run and are not fit to benchmark scores.

The training corpus is split by source video rather than by story. A single source video can yield several stories, so a story-level split would place stories derived from the same video on both sides of the split; splitting by source video keeps every story derived from one video on one side.

C.4 Prompt alignment

We retain the benchmark's scene, camera, global-action, and local-action rubrics and story-level aggregation, including integer rounding for local actions. PA averages the four scores, scaled from 0–4 to 0–100. Local actions use matched character crops.

The four-component results appear in Table 2 in Section 5.2.

Published comparison sources. The published panel of Table 1 reproduces the upper panel of Table 2 in Zhuang et al. (2026)'s CVPR 2026 main paper (PDF page 7; proceedings page 9461), cross-checked against preprint v5. Published values retain the source precision and were obtained under the benchmark authors' inference settings; the lower panel and supplementary Table S4 use a different split and supply no values. PA and CP appear only in local comparisons (Tables 2 and 3), because published and local evaluations differ in PA configuration and CP scoring-code version.

D Qualitative Comparisons and Controller Trace

D.1 Qualitative cases

Figures 2 and 3 compare original BAGEL, single-pass ASM+SIG, and feedback-controlled StoryRelay, with ground-truth (GT) keyframes as visual anchors. These curated, independent rollouts show output differences, not direct edits of ASM+SIG images; the within-rollout retry in Appendix D.3 separately documents a controller correction.

Switching active characters. In Figure 2, ASM+SIG carries the striped-shirt patient into Shot 3 in place of the requested older doctor: the count is correct, but the cast is not. StoryRelay shows the requested doctor beside `Subject3`, without the carried-over patient. Its displayed attempts are 1/2/1, so the Shot 3 difference is not itself a rejection-and-retry example.

Separating co-occurring identities. In Figure 3, ASM+SIG duplicates the red-clad patient in Shot 2 and adds an extra patient in Shot 4. StoryRelay shows one red-clad and one green-clad patient in both views, preserving identity and count together. All four displayed StoryRelay outputs are first attempts, rather than corrected versions of the ASM+SIG candidates.

D.2 ADDITIONAL CASES

Figures 6–8 use the same comparison format. Recorded shot texts match across methods except for terminal punctuation; these examples are separate from the ViStoryBench quantitative evaluation.

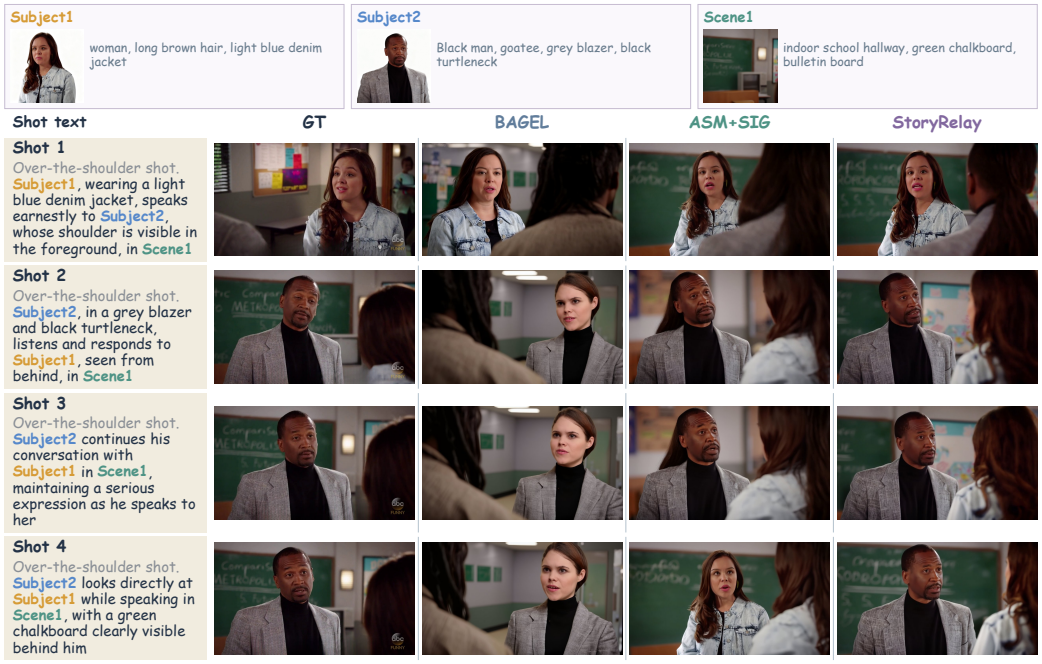


Figure 6: Dialogue identity across views. In Shot 4, ASM+SIG replaces the requested camera-facing man with the denim-jacket woman. StoryRelay retains the man's face, grey blazer, and black turtleneck, with the woman only partially visible in the foreground.

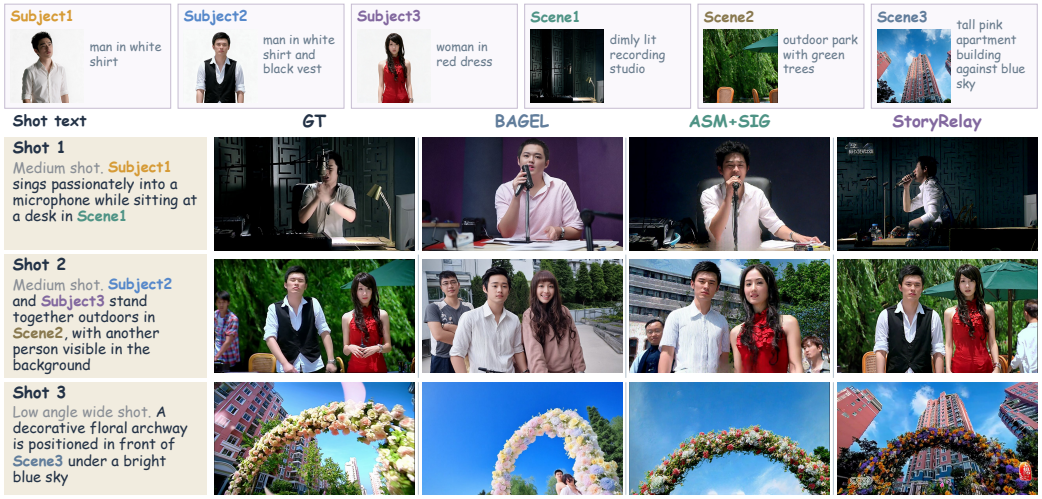


Figure 7: Character and scene changes. ASM+SIG omits the man's black vest and the tree-and-parasol setting in Shot 2, and the pink buildings in Shot 3; StoryRelay retains these reference details. This comparison concerns visible clothing and scene context, not exact sitting-pose agreement in Shot 1.



Figure 8: Identity retention with a composition difference. Both ASM+SIG and StoryRelay preserve the returning character and the two intervening identities. StoryRelay additionally matches the GT pair's left–right arrangement in Shot 2; this is not evidence of an identity correction.

D.3 Controller trace

Table 8 records all five attempts from a four-shot rollout using two subject and two scene reference images cropped from an illustrative asset. Shot 4 requests the woman alone in the greenhouse.

Table 8: Four-shot controller trace (one-based IDs). Original references remain in every attempt; "Final accepted?" marks the chosen, checker-passing output.

Shot	Selected history	Attempt	Checker verdict	Correction	Final accepted?
1	None	1	Pass	None	Yes
2	None	1	Pass	None	Yes
3	Shot 2	1	Pass	None	Yes
4	Shot 2	1	Reference failure	Drop Shot 2; regenerate	No
4	None	2	Pass	None; append to history	Yes



Attempt 1: Shot 2 history; rejected

Attempt 2: no history; accepted

Figure 9: Shot 4 retry from the same rollout. Dropping Shot 2 and regenerating removes the extra man. The accepted image still does not clearly depict the requested watering action.

The failed verdict's local `Scene2` denotes global `Scene1` (greenhouse). Only attempt 2 enters memory; pruning and resampling are not isolated in this example.