Cinematic Compiler: A Typed Intermediate Representation with a Deterministic Pre-Generation Constraint Checker for Multi-Shot Keyframe Continuity

Akash Sagi
Florida State University
Tallahassee, Florida, USA
ss24bc@fsu.edu

Karthik Reddy Vemireddy
Florida State University
Tallahassee, Florida, USA
kv23b@fsu.edu

Sai Ganesh Akula
Florida State University
Tallahassee, Florida, USA
sa23k@fsu.edu

Abstract

LLM-planned generation pipelines compile a natural-language brief into an intermediate representation before invoking a downstream generator, the same pattern used for structured programs (VisProg, ViperGPT, LayoutGPT) and agent pipelines (MetaGPT, DSPy). That plan is typically checked, if at all, only after generation. We introduce a **deterministic constraint pass that verifies the compiled plan before any generation spend**. Applied to multi-shot video, where text-to-video models lose cross-shot continuity as a character's face, costume, or a recurring prop drifts between cuts, Cinematic Compiler compiles a story brief into a typed Cinematic IR, checks every shot prompt against the continuity facts its entities require *before any pixels are generated*, and anchors shot generation on a character reference image at inference time, on frozen, commodity backends with no training. Three results: **(1)** the compiled IR satisfies 100% of per-shot continuity constraints versus 2% for direct prompting and 23% for LLM-expanded prompting (a true-by-construction fact-propagation guarantee, distinct from pixel-level satisfaction, which we measure separately), and the checker rejects both seeded impossible briefs before spend; **(2)** reference anchoring raises cross-shot identity similarity from 0.605 to 0.714 (12/13 briefs, Wilcoxon p=0.003, d_z=0.91), where matched text-only prompting does not (Δ=+0.027, n.s.) — a same-backbone control shows most of this gain is attributable to the reference-conditioned backend rather than conditioning alone; **(3)** reference-anchored repair raises identity-vs-reference $0.628 \rightarrow 0.713$ over 46 audit-failing shots ($p=1\times10^{-4}$), where text-prompt repair does not move identity. We also show the field's default continuity metric, pairwise cross-shot embedding similarity, can be gamed by degenerate near-static output, and report identity-vs-reference jointly with a dynamic-degree score to guard against it. All quantitative results are keyframe-level; we additionally validate generalization on briefs from an uninvolved external contributor (exposing and fixing a real conflict-detector gap) and on independently authored ViStoryBench stories. The demo lets attendees compile, check, edit, and repair a short film interactively.

1 Introduction

Single-clip text-to-video quality is no longer the bottleneck for narrative generation: frontier models exceed 93% VBench subject consistency (Huang et al., 2024), yet a five-shot film made from five independent generations shows a different protagonist in every cut.[1] EntityBench quantifies the failure: cross-shot entity consistency degrades sharply with recurrence distance (He et al., 2026). The 2025–26 response is a wave of multi-agent film systems (Wu et al., 2025; Huang et al., 2026; Shi et al., 2025; Zhou et al., 2026a) and generator-side mechanisms: entity memory (Wei et al., 2026; Zhou et al., 2026b), entity-grounded shot scheduling (Lai et al., 2026), storyboard anchoring (Zhang et al., 2026b), and closed-loop multimodal-feedback enforcement (Sun et al., 2026). All of these verify output *after* generation, and most require model access or training.

Cinematic Compiler instead treats **continuity as a checked property of a symbolic plan, verified before generation, on frozen commodity backends** (Fig. 1). The system compiles a one-line brief into a typed Cinematic IR spanning narrative, shots with per-shot `must_preserve` facts, ten design departments, and a continuity graph of characters, props, and locations with appearance schedules. Two deterministic passes run before any generation spend: a constraint-satisfaction pass and a conflict detector that rejects impossible briefs outright. At generation time, shots are anchored on the compiled character reference image via commodity im-

[1] Code:https://github.com/ssrakash5/cinematicompiler. Live demo: http://8.222.231.140/.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 17–24
November 6-10, 2026 ©2026 Association for Computational Linguistics

age editing, and a vision-model audit scores each shot against the graph.

We make four measured contributions. **(1) Pre-generation constraint checking.** The deterministic pass reaches 100% per-shot constraint satisfaction versus 2% for direct prompting and 23% for LLM-expanded prompting; an ablation shows the pass itself, not merely having an IR, supplies the majority of the gain (36.5% with the pass disabled). This figure is true by construction — fact propagation into the prompt, not pixel fidelity, which §5 measures separately. The conflict detector classifies all 15 benchmark briefs correctly. **(2) A controlled decomposition of where cross-shot identity lives.** Holding prompts and references fixed and swapping only the generation mechanism, text-level identity constraints do not transfer to pixels (Δ=+0.027, n.s.), while reference anchoring raises identity-vs-reference similarity 0.605→0.714 (+0.108±0.119, 12/13 briefs, Wilcoxon p=0.0034, d_z=0.91); the gain anti-correlates with the text-only baseline (Pearson r=−0.810, p<0.001). **(3) A gaming-resistant, human-validated continuity evaluation.** We exhibit an adversarial case where pairwise cross-shot embedding similarity scores a degenerate near-duplicate baseline *above* a visibly better film (0.86 vs. 0.73), and define the corrected protocol. Three independent raters scoring 90 shots stratified across the metric's full range correlate with it at Pearson r=0.82 (Spearman ρ=0.80, both p<0.001), with substantial inter-rater agreement (Krippendorff's α=0.72). **(4) CineBench-Plan v0 and a reusable harness**: 22 briefs (15 team-authored + 7 external) with gold constraints and an offline metric suite that re-scores stored artifacts with no further generation calls.

2 Related Work

Multi-agent film generation. MovieAgent (Wu et al., 2025) plans films with hierarchical CoT agents; ViMax (Huang et al., 2026), AniMaker (Shi et al., 2025), and DramaDirector (Zhou et al., 2026a) extend the recipe; FilmAgent (Xu et al., 2025) stages crews in Unity 3D; Anim-Director (Li et al., 2024) self-refines an animation pipeline. Every critic in this family judges *generated output*. We introduce a deterministic symbolic check that runs before generation and rejects impossible plans before any spend.

Cross-shot consistency mechanisms. GroundShot (Lai et al., 2026) schedules shots over grounded entities; STAGE (Zhang et al., 2026b) anchors on storyboards; Memento (Wei et al., 2026) and EntityBench's EntityMem maintain entity memory; ARGUS (Meng et al., 2026) and VideoGen-of-Thought (Zheng et al., 2024) inject identity-preserving conditioning into the generator. These operate inside or alongside the generator. Our contribution is the controlled decomposition: measuring text-level and image-level continuity separately under matched prompts, finding the gain concentrates where prompting fails.

Verify-before-render (2026 concurrent work). MUSE (Sun et al., 2026) frames story generation as closed-loop constraint enforcement, but its loop verifies *generated* multimodal output via feedback — the same shape as our own audit→repair — and reports no deterministic satisfaction rate. GEST (Cudlenco et al., 2026) compiles a story into a structured event graph executed deterministically inside a 3D game engine, guaranteeing physical validity by construction, but on a different backend class (a deterministic renderer, not a frozen generative model) and target (physics, not entity continuity). Our checker runs *before* generation, against a frozen commodity backend, and reports the pass's isolated contribution directly (§5.1); neither reports this combination.

Benchmarks. VBench (Huang et al., 2024) and EvalCrafter (Liu et al., 2024) measure single-clip quality; EntityBench (He et al., 2026) and MuSS (Zhang et al., 2026a) measure cross-shot consistency of generated output. CineBench-Plan evaluates the *plan* before pixels exist. SketchVerify (Huang et al., 2025) shares our verify-before-synthesis motivation for physics plans; our checker verifies symbolic continuity constraints.

LLM planning and structured IRs. VideoDirectorGPT (Lin et al., 2024) established LLM video plans; LayoutGPT (Feng et al., 2023) and VisProg (Gupta and Kembhavi, 2023) established LLM-emitted structured programs; DSPy (Khattab et al., 2024) compiles LM pipelines against a metric; MetaGPT (Hong et al., 2024) passes structured documents between role agents. We specialize the paradigm to a typed, temporally-indexed cinematic IR whose value the checker ablation demonstrates.

3 System

Cinematic IR. A job compiles through six stages: script, storyboard, design bible (ten departments),

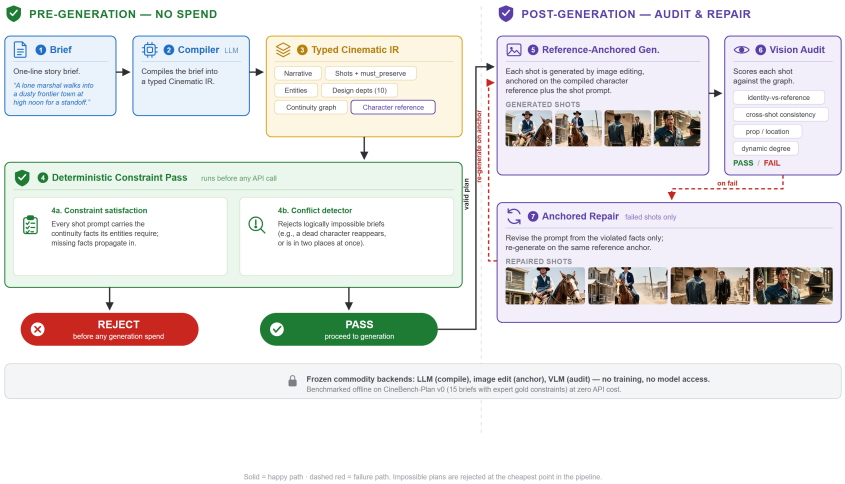


Figure 1: Brief → typed IR → deterministic constraint pass and conflict detector (pre-generation, no spend) → reference-anchored generation on a frozen backend → vision audit and anchored repair (post-generation). Impossible briefs are rejected before any generation cost; audit-failing shots re-generate on the same reference anchor rather than a text re-roll. Generated and repaired shots shown are real pipeline output.

System	Verifies pre-gen.?	Frozen backend?	Reports CSR?	Continuity mechanism
Prior multi-agent film & cross-shot systems†	✗	varies	✗	LLM/VLM critic, scheduling, or entity memory (post-hoc)
MUSE (Sun et al., 2026)	✗ (verifies output in loop)	unspecified	✗ (human-validated)	closed-loop multimodal feedback
GEST (Cudlenco et al., 2026)	✓ (graph checked pre-exec)	✗ (3D game engine)	validity-by-construction	simulator constraints
Cinematic Compiler	✓	✓	✓ **(100% vs. 2%/23%)***	deterministic checker + ref. anchoring

Table 1: Comparison with existing systems (§6.2). CSR = constraint-satisfaction rate. †MovieAgent, ViMax, AniMaker, DramaDirector, FilmAgent, Anim-Director, GroundShot, STAGE, Memento, VideoMemory, EntityMem. *Plan-level, true by construction (§3); pixel-level satisfaction is measured separately (§§5.3–5.4).

continuity graph, references, shots. Shots carry `must_preserve` fact lists; the graph tracks characters, props, and locations with `appears_in` schedules and identity/costume attributes. The same schema represents benchmark gold constraints and live pipeline output, so one metric path scores both (spec released as `IR_SPEC.md`).

Pass 1: constraint satisfaction (pre-generation, deterministic). For each shot, the union of its `must_preserve` facts, the attributes of characters scheduled to appear, the required props, and location facts must appear in the compiled shot prompt; missing facts are propagated in. Disabling propagation drops satisfaction from 100% to 36.5%. This checks fact propagation into text, not pixel fidelity, which Pass 4 and §§5.3–5.4 assess empirically.

Pass 2: conflict detection (pre-generation, deterministic). Targeted contradiction checks (weather vs. locked location; costume vs. locked wardrobe with no change beat) reject impossible briefs before spend; 15/15 briefs classified correctly.

Pass 3: reference anchoring (generation-time). Shot images are generated by identity-preserving image editing conditioned on the compiled character reference portrait plus the compiled shot prompt, on a frozen commodity endpoint.

Pass 4: vision audit and anchored repair. A VLM scores each shot against its constraint checklist (threshold 0.75); failing shots regenerate through Pass 3 with the repair prompt. Text-prompt repair does not move identity (§5.4); reference-anchored regeneration through Pass 3 does, and is the repair path we report.

Harness. A model-fallback client (text, vision, t2i, image-edit chains) completed all reported batches with zero generation errors; artifacts re-score offline with no further calls.

4 Evaluation Methodology

Pairwise similarity is gameable. The default cross-shot metric, mean pairwise CLIP/DINO similarity, rewards monotony: a baseline emitting near-duplicate generic frames scores *higher* (0.86) than a visibly better, more varied film (0.73) on the same brief. Any planning layer that increases shot variety is penalized by construction.

Metric of record. We score (a) *identity-vs-reference*: mean CLIP similarity of each shot to the

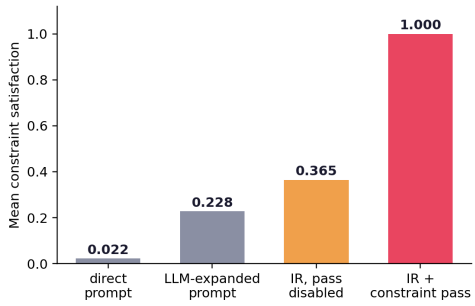


Figure 2: Per-shot continuity-constraint satisfaction across prompting conditions (15 briefs).

film's character reference portrait, anchoring the comparison to a fixed identity; and (b) *dynamic degree*: mean pairwise visual change across shots, reported alongside every continuity number so static output is visible.

Human validation. Since CLIP similarity can be brittle, three raters independently scored 90 shots (1–5 Likert, same-character confidence), stratified across the metric's full score range (0.345–0.958). Mean human rating correlates with identity-vs-reference at Pearson r=0.82 (p<0.001) and Spearman ρ=0.80 (p<0.001); inter-rater agreement is substantial (Krippendorff's α=0.72, 3 raters). Sampling, protocol, and analysis are released.

5 Results

5.1 Constraint adherence and pass ablation

Fig. 2: direct prompting satisfies 2.2% of per-shot continuity constraints, LLM-expanded prompting 22.8%, the IR with the constraint pass disabled 36.5%, and the full pass 100% (15 briefs, offline, deterministic). Conflict detection: 15/15. The increase from 36.5% to 100% indicates that constraint propagation, rather than the IR alone, accounts for most of the adherence gain.

5.2 External benchmark: a generalization gap we found and fixed

The 15 briefs above are team-authored, risking rules tuned to our own test cases. A contributor uninvolved in the design, given only the schema, added 7 briefs: 5 consistent, plus 2 impossible with new contradiction types — a prop destroyed then required intact later, and a fixed identity marker required on the opposite side. This exposed a real gap: our original two-rule detector (weather, costume) scored 3/7 (43%) — 2 false positives from checking a shared `must_preserve` list against each present character's lock individually rather than the union of everyone present, and 2 false negatives, since neither new type had a matching rule. We fixed the union bug and added identity-orientation and prop-state rules; re-run: external detection **7/7**, zero regression on the original 15 (combined 22-brief benchmark: A=0.018, C=0.227, E0=0.369, E/F=1.000, conflict detection 22/22). We report the pre-fix 3/7 alongside the post-fix 7/7, since the gap it exposed is itself informative.

5.3 Where cross-shot identity lives

Text-level constraints do not transfer to pixels: across 6 briefs with matched references, identity-vs-reference for direct prompting vs. the full text-only compiler is 0.533 vs. 0.560 (Δ=+0.027, 3/3 per-brief split, n.s.). The compiler's prompts carry the identity facts (§5.1); the frozen backend does not preserve them. Reference anchoring closes the gap (Fig. 3, Fig. 4a): holding the reference and compiled prompts fixed and swapping only the generation mechanism, identity-vs-reference rises 0.605 $\rightarrow$ 0.714 (+0.108±0.119; Wilcoxon p=0.0034; d_z=0.91; 12/13 briefs). The single loss is the brief whose text-only baseline was already highest (0.719), and across briefs the gain anti-correlates with the text-only baseline (Pearson r=−0.810, p=0.0008): anchoring contributes most exactly where prompt-level continuity fails worst (e.g., 0.445 $\rightarrow$ 0.755; 0.489 $\rightarrow$ 0.844). *Trade-off:* anchoring reduces dynamic degree 0.267 $\rightarrow$ 0.188 (p=0.0002): anchoring identity reduces shot variety, and we report both numbers in every table so this trade-off stays visible.

Same-backbone control (n=6). The comparison above also swaps backbones (wan2.1-t2i-turbo $\rightarrow$ qwen-image-edit-max), confounding it with conditioning. Holding the backbone fixed and only toggling `input.ref_img` on wan2.1-t2i-turbo (fresh reference, same prompts), identity-vs-reference rises 0.623 $\rightarrow$ 0.646 (Δ=+0.023±0.034, 5/6 briefs, paired t p=0.165, Wilcoxon p=0.156, d_z=0.66) — real and same-direction, but about a fifth of the +0.108 backbone-swap gain, n.s. at this n. Most of the reported gain is attributable to the edit-capable backend, with a modest additional contribution from conditioning alone.

5.4 Repair

Text-prompt repair (rewriting the shot prompt from audit findings and regenerating via text-to-image)

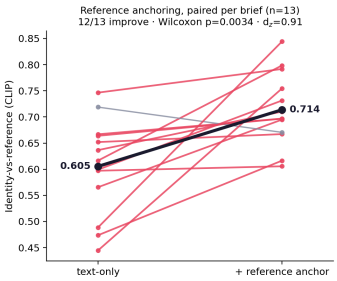


Figure 3: Paired identity-vs-reference per brief, text-only vs. reference-anchored generation under identical compiled prompts (n=13).

does not move identity (audit $0.630 \rightarrow 0.635$ over 6 briefs, clearing the threshold on none): repair that stays in prompt space cannot correct pixel-level identity drift. *Reference-anchored repair does.* Regenerating all 46 audit-failing shots across the 13 briefs through Pass 3 (image edit on the character reference with the stored repair prompt) raises identity-vs-reference $0.628 \rightarrow 0.713$ (38/46 shots improved, Wilcoxon $p{=}1{\times}10^{-4}$, $d_z{=}0.62$), with zero generation errors. The VLM audit score moves only $0.599 \rightarrow 0.612$ (n.s.): the deterministic identity metric registers a repair effect the noisy vision audit largely misses: evidence that verification should weight deterministic identity measurement over VLM scoring, consistent with §4.

5.5 Reproducibility: a negative result

Under default sampling the compiled plan is *not* reproducible: across 5 reruns of each of 3 briefs, mean structural IR distance (normalized edit distance over shot-level structural fields; 0 = identical, 1 = disjoint) is 0.77 (per-brief 0.68–0.83). The pipeline invents continuity facts from the one-line logline, so each run invents a different set, and no seed is plumbed into the text calls; this motivates gold-constraint ingestion and seed control (§8) and does not affect §§5–5.4, which hold the plan fixed per condition.

5.6 System

IR validity 1.0 on gold briefs. All live batches completed with zero generation errors. Wall-clock latency across the 13 benchmark films (script through shot generation, harness timestamps) is 20.3 minutes mean / 25.9 minutes median (range 6.8–35.9 minutes, driven by DashScope rate limiting rather than compute). Pinned models: qwen3.5-flash (text), qwen-vl-plus (audit), wan2.1-t2i-turbo (t2i), qwen-image-edit-max (anchoring).

6 Demonstration

Attendees drive the pipeline from a canvas UI[2]: **(1) Compile** a one-line brief and inspect the IR building stage by stage. **(2) Check**: run the constraint pass on a seeded conflict brief and watch it rejected before generation with the contradiction named; post-hoc critics only evaluate output after generation. **(3) Break and steer**: chat with the Design Bible node ("change her coat to red"); the edit propagates through the graph into references and every scheduled shot, downstream stages regenerate, and the checker re-verifies — a live, editable plan, not a one-shot LLM output. **(4) Extend**: attendees drop a custom agent into any stage with a one-line system prompt; its output slots into the graph as a new node. **(5) Generate and audit**: shots render reference-anchored; the vision audit scores each; the film assembles. If connectivity fails, every benchmark brief has stored jobs that replay through the same UI. Fig. 4b shows the Shot 01 detail panel from a stored job.

6.1 Target Audience

Two audiences: (1) researchers building LLM-planned generation pipelines, who need a worked example of verifying a compiled plan before spending generation budget on it, and can reuse the IR, checker, and custom-agent interface outside the video domain; (2) narrative content creators and tooling teams evaluating multi-shot AI video systems, who get a checked, editable alternative to blind re-prompting.

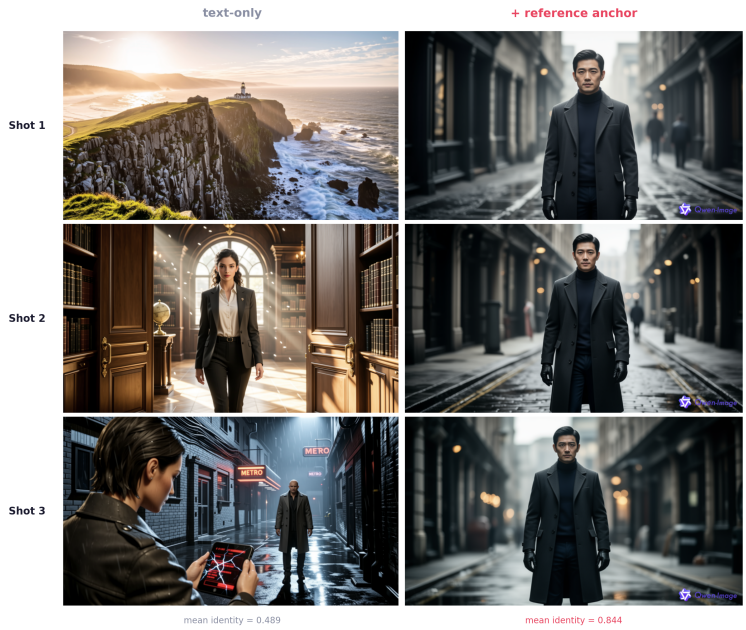
6.2 Comparison with Existing Systems

Table 1 (§2) situates Cinematic Compiler against the closest 2026 systems: MUSE verifies generated output in a feedback loop; GEST verifies a structured plan inside a deterministic 3D game engine rather than a generative backend. §2 details each.

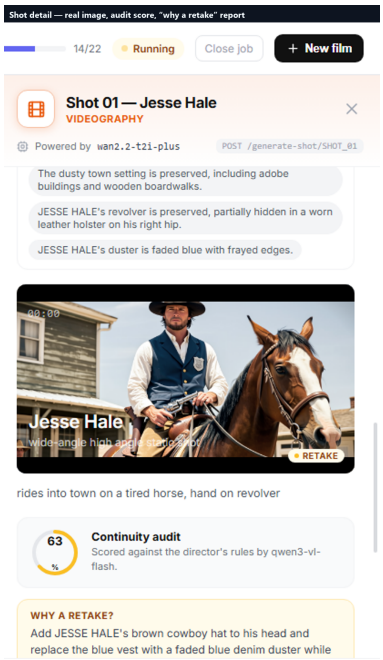
7 Limitations

Scope: keyframe continuity, not motion/temporal coherence. The system assembles a final video via an image-to-video backend and concatenation (§3), and the demo plays a full film, but every quantitative result (§5) is measured on the

[2]Video demo: https://drive.google.com/file/d/1yrI7hZjjy863PW6OzdCTZ6WH1kYty4Ur/view?usp=sharing.



(a) Same compiled prompts per row, only the generation mechanism swapped. Left column: text-only generation loses the subject entirely by Shot 2. Right column: reference-anchored generation holds identity across all three shots.



(b) Shot 01 detail panel from the live canvas UI against a stored job (read-only, zero new generation): the audit score ring (63%, below the 0.75 threshold) and the "why a retake" report naming exactly which `must_preserve` facts the shot failed.

Figure 4: Reference-anchored generation vs. a text-only baseline (a), and the live audit/retake UI that drives repair (b).

compiled shot images, before the image-to-video step; motion quality, intra-clip temporal coherence, and transition continuity are inherited from the frozen backend and unevaluated here. Read this paper's claims as claims about **cross-shot keyframe continuity**; full-film motion/transition evaluation is future work.

Identity-vs-reference uses CLIP embeddings; we validate it against human judgment (§4, r=0.82, n=90), but on a stratified sample, and the adversarial pairwise case is one brief. The benchmark is 15 team-authored briefs plus 7 external ones (§5.2, uninvolved contributor, 2 new contradiction types) — this exposed and let us fix a real generalization gap in the conflict detector rather than confirming a clean result. A larger externally-sourced benchmark at $\geq$100 briefs remains v1 future work; 7 briefs bounds rather than eliminates the risk that our taxonomy misses classes a larger benchmark would surface. As a supporting check on genuinely third-party content, we replicated §5.3 on 9 ViStoryBench (Zhuang et al., 2025) stories (real characters/shots/photos, our own metric, not their GPU-only evaluator): 0.635 $\rightarrow$ 0.702 (Δ=+0.066$\pm$0.093, 6/9 improved, paired t p=0.067, Wilcoxon p=0.129, d_z=0.71) — same direction and comparable effect size to the main result, n.s. at this n; a full run against their official evaluator is future work. The pipeline invents continuity facts from the logline rather than ingesting gold constraints, so gold-referenced satisfaction for live runs understates the mechanism of §5.1. Reference anchoring uses a single character portrait per film; multi-character anchoring and the dynamic-degree trade-off are open. §5.3's comparison also changes backend (text-to-image vs. image-editing); the same-backbone control there isolates conditioning. Pixel quality is inherited from the backends; we make no pixel-quality claims.

8 Conclusion

Continuity is better treated as a checked property of a plan plus an anchored property of generation than an emergent hope of a generator: a deterministic pre-generation pass reaches perfect adherence where prompting reaches 2%, and a reference anchor delivers pixel-level identity prompting did not preserve.

References

Nicolae Cudlenco, Mihai Masala, and Marius Leordeanu. 2026. Agentic video generation: From text to executable event graphs via tool-constrained llm planning. *arXiv preprint arXiv:2604.10383*.

Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Layoutgpt: Compositional visual planning and generation with large language models. In *NeurIPS*.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *CVPR*.

Ruozhen He, Meng Wei, Ziyan Yang, and Vicente Ordonez. 2026. Entitybench: Towards entity-consistent long-range multi-shot video generation. *arXiv preprint arXiv:2605.15199*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *ICLR*.

Lingxuan Huang, Sizhe He, Hengji Zhou, Liqiang Nie, Lianghao Xia, and Chao Huang. 2026. Vimax: Agentic video generation. *arXiv preprint arXiv:2606.07649*.

Yidong Huang, Zun Wang, Han Lin, Dong-Ki Kim, Shayegan Omidshafiei, Jaehong Yoon, Yue Zhang, and Mohit Bansal. 2025. Planning with sketch-guided verification for physics-aware video generation. *arXiv preprint arXiv:2511.17450*.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines. In *ICLR*.

Yixuan Lai, Tianjia Shao, Kun Zhou, Weijia Dou, Siyu Zhu, and Jingdong Wang. 2026. Groundshot: Visually consistent multi-shot long video generation via entity-grounded shot scheduling. *arXiv preprint arXiv:2606.20799*.

Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, and Min Zhang. 2024. Anim-director: A large multimodal model powered agent for controllable animation video generation. In *SIGGRAPH Asia*. ArXiv:2408.09787.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. 2024. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. In *COLM*.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In *CVPR*.

Zijie Meng, Jiwen Liu, Yufei Liu, Chengzhuo Tong, Xiaoqiang Liu, Yuanxing Zhang, Yulong Xu, and Pengfei Wan. 2026. Argus: Stacked multi-view identity mosaic injection for subject-preserving video generation. *arXiv preprint arXiv:2606.11670*.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. 2025. Animaker: Multi-agent animated storytelling with mcts-driven clip generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. ArXiv:2506.10540.

Wenzhang Sun, Zhenyu Wang, Zhangchi Hu, Chunfeng Wang, Hao Li, and Wei Chen. 2026. Muse: A multi-agent framework for unconstrained story envisioning via closed-loop cognitive orchestration. *arXiv preprint arXiv:2602.03028*.

Xuan Wei, Longbin Ji, Guan Wang, Xiangrui Liu, Zhenyu Zhang, Shuohuan Wang, Yu Sun, and Qingqi Hong. 2026. Memento: Reconstruct to remember for consistent long video generation. *arXiv preprint arXiv:2606.14667*.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. 2025. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*.

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. 2025. Filmagent: A multi-agent framework for end-to-end film automation in virtual 3d spaces. *arXiv preprint arXiv:2501.12909*.

Haojie Zhang, Di Wu, Bingyan Liu, Linjie Zhong, Yuancheng Wei, Xingsong Ye, Nanqing Liu, and Yaling Liang. 2026a. Muss: A large-scale dataset and cinematic narrative benchmark for multi-shot subject-to-video generation. *arXiv preprint arXiv:2604.23789*.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. 2026b. Stage: Storyboard-anchored generation for cinematic multi-shot narrative. In *CVPR*. ArXiv:2512.12372.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. 2024. Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention. *arXiv preprint arXiv:2412.02259*.

Hengji Zhou, Sijie Liu, Jianrun Chen, Xingchen Zou, Lianghao Xia, and Liqiang Nie. 2026a. Dramadirector: Geometry-guided short drama generation. *arXiv preprint arXiv:2606.24107*.

Jinsong Zhou, Yihua Du, Xinli Xu, Luozhou Wang, Zijie Zhuang, Yehang Zhang, Shuaibo Li, Xiaojun Hu, Bolan Su, and Ying-cong Chen. 2026b. Videomemory: Toward consistent video generation via memory integration. *arXiv preprint arXiv:2601.03655*.

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. 2025. Vistorybench: Comprehensive benchmark suite for story visualization. *arXiv preprint arXiv:2505.24862*.