CANVAS: PLAN THE SHOOT, NOT JUST THE STORY, FOR LONG VIDEO

Anonymous authors
Paper under double-blind review

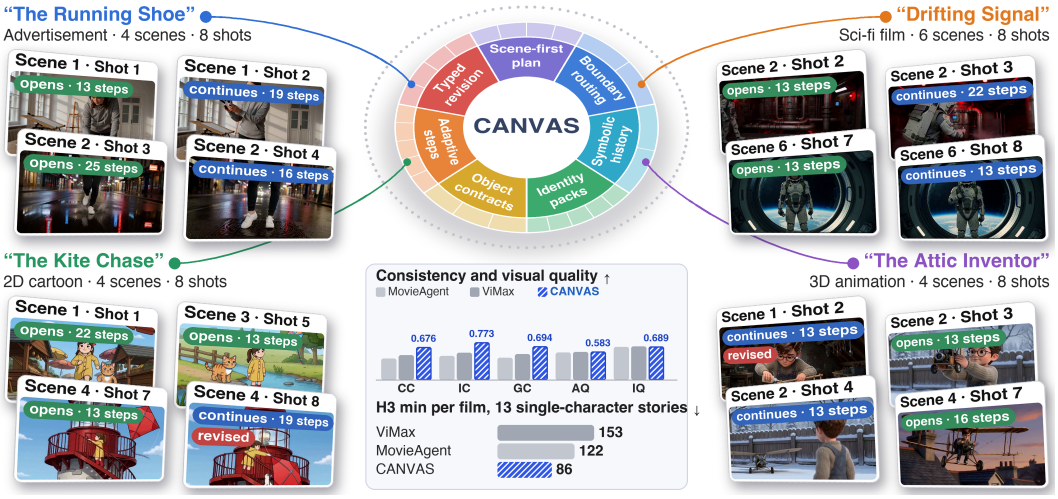


Figure 1: **CANVAS plans each shot before rendering.** Four selected ViMax-Bench films generated with the Qwen3.8-27B planner and MiniMax-H3 video generator (first two shots of scenes 1 and 2). Each card shows whether the shot opens from identity references or continues from the previous shot's last frame, its denoising steps, and any typed revision. Center: consistency, VBench aesthetic quality (AQ), and imaging quality (IQ) across 35 stories; MiniMax-H3 inference time per second of video across 20 stories on RTX 6000 Ada GPUs. The legend specifies step settings (Appendix G).

ABSTRACT

Video generators synthesize striking short clips, but assembling them into coherent multi-shot films remains difficult. Existing approaches plan the story and shots, then render each shot with the same conditioning policy and compute budget. We argue that coherence hinges on two shot-level decisions: what visual context to inherit and how much computation to spend. We introduce CANVAS, a training-free framework that makes both decisions before generation. It compiles a story into an executable shooting plan: pixels carry over only within continuous action, while structured state carries identities, objects, and narrative history across cuts. A film-level budget assigns computation according to shot difficulty, and a verifier directs revisions to specific failures. On ViMax-Bench, with the language model, image generator, and video generator held fixed, CANVAS raises global consistency from 0.528 for ViMax and 0.452 for MovieAgent to 0.694, with VBench aesthetic and imaging quality on par with ViMax. It spends 0.76 times the denoising of a uniform 25-step pass on the same 35 stories; generation takes 136 minutes per film, versus 129 for ViMax and MovieAgent. The consistency gains are supported by three evaluator families and a human study, and persist without frame chaining. We will release the code to the public.

1 INTRODUCTION

Video generators synthesize short clips with realistic texture, lighting, and motion (Blattmann et al., 2023a; Yang et al., 2024; Kong et al., 2024; Team Wan et al., 2025). Agentic systems assemble them

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.

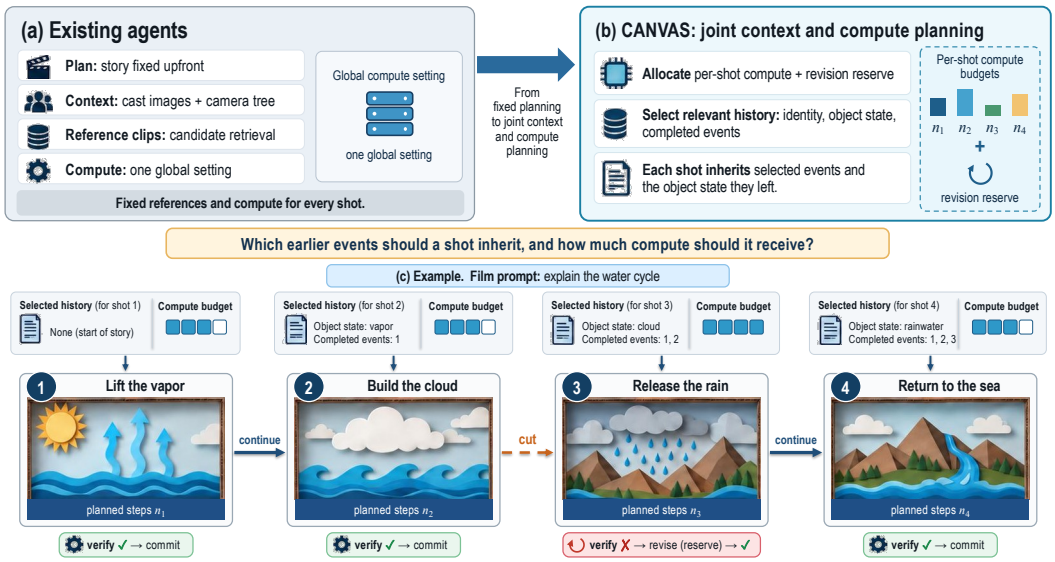


Figure 2: **What each shot inherits and how much compute it receives.** (a) MovieAgent and ViMax plan narratives and references; in ViMax, each shot draws on a fixed pool of references (portraits of visible characters and a parent camera frame) under one global generation setting. (b) CANVAS selects prior events by relevance and assigns each shot an allowance n_i within a film-level budget, reserving capacity for revision. (c) *Continue*: the shot starts from the previous shot's last frame; *cut*: it opens a new scene without that frame, and only the selected history crosses.

into multi-shot films by planning a story, specifying shots, and invoking a generator for each one (Lin et al., 2023a; Wu et al., 2025; Huang et al., 2026a). Yet a coherent script does not guarantee a coherent film. Continuing from the wrong frame can repeat a completed action; starting from scratch can change a character or setting; carrying a frame across a scene cut can obstruct the scripted transition. The issue is how each shot is executed: which information should it inherit?

A second decision is how much computation to allocate. Pipelines commonly use the same generation settings for a static establishing shot and a complex interaction, such as two people passing a wet clay bowl. Extra denoising cannot recover an object missing from the prompt or identify an intended scene cut. Both context and computation must therefore be decided from the shot dependencies before rendering.

Existing systems largely leave these decisions implicit. Hierarchical agents divide stories into scenes and shots and assign production roles (Wu et al., 2025; He et al., 2024a; Xu et al., 2025). ViMax retrieves narrative context and uses camera-dependent reference frames (Huang et al., 2026a); multi-shot and storyboard models improve consistency within the generator (Zhou et al., 2024; Zhang et al., 2025b; Meng et al., 2025). Yet agentic pipelines typically apply fixed context rules and global generation settings across shots (Figure 2). We therefore ask: *Can an authored plan determine both the visual inputs and the generation budget of every shot before rendering?*

CANVAS answers this question with an executable shooting plan (Figure 3). Its central distinction is between pixels and story state. Each shot either continues from the previous shot's final frame or opens from identity references, while structured records carry identities, objects, and narrative history across both types of boundary. A shot-level difficulty estimate then assigns a denoising tier within a fixed film budget that reserves capacity for revisions. The plan records each committed take's inputs, denoising steps, and verification results for reproducibility.

Three checks make the plan executable. Object hand-offs are validated before rendering to keep story state aligned with visual inputs. Budget allocation ensures that harder shots receive at least as many steps as easier ones. After rendering, a verifier targets revisions to specific failures and flags shots it cannot repair. On ViMax-Bench, CANVAS raises global consistency from 0.528 for ViMax and 0.452 for MovieAgent to 0.694, with VBench aesthetic and imaging quality on par with ViMax. The gain persists when all continuation links are removed. Figure 1 compares CANVAS

with 13–25-step tiers against 25-step baselines, showing an advantage even at a lower step budget. The contributions of this work are:

- **Per-shot context and compute planning.** We introduce CANVAS, a training-free framework that compiles a story into an executable shooting plan. Before rendering, each shot specifies its boundary, relevant prior events, object-action contract, and denoising tier.
- **Boundary-aware context routing and difficulty-aware compute allocation.** Pixels pass only across continuations, limiting each shot to one inherited frame; structured records carry identity, objects, and story history across cuts. Difficulty-based denoising tiers respect a film-level budget, and a typed verifier directs revisions to failed shots.
- **Improved consistency.** CANVAS surpasses ViMax and MovieAgent in cross-shot and cross-scene consistency with comparable visual quality, using fewer denoising steps per shot (16.4 on average against 25). The gains hold across three evaluator families and a blinded human evaluation. Ablations show that frame chaining supports within-scene continuity, while identity references sustain character consistency across scene cuts.

2 RELATED WORK

Long video generation. Video diffusion and transformer backbones synthesize short clips from text and visual conditions (Ho et al., 2022b; Blattmann et al., 2023a; Yang et al., 2024; Kong et al., 2024; Team Wan et al., 2025). Streaming, hierarchical, and memory mechanisms extend generation over time (Yin et al., 2023b; Henschel et al., 2025; Lu et al., 2024; Zhou et al., 2026), while cross-image attention, storyboards, and multi-shot models impose sequence structure (Zhou et al., 2024; Zhang et al., 2025b; Meng et al., 2025). Earlier approaches generate autoregressively over tokens or latents (Villegas et al., 2022; Ge et al., 2022; Harvey et al., 2022); story-level generators condition shots on memories or keyframes from earlier shots (Zhang et al., 2025a; Zhao et al., 2024; Xiao et al., 2025). ContextMaster routes shot history within a fixed context budget in a trained model (Guo et al., 2026), and CineWeaver reshapes a pretrained model's attention to separate shots (Huang et al., 2026b). CANVAS operates at the shot level: it keeps the generator frozen and selects each shot's conditioning and generation settings.

Agentic film production. VideoDirectorGPT plans scenes and entities (Lin et al., 2023a); MovieAgent assigns hierarchical production roles (Wu et al., 2025); and other systems explore collaborative or cinematic planning (He et al., 2024a; Xu et al., 2025; Huang et al., 2025; Wang et al., 2026). MovieFactory, Animate-A-Story, and VideoGen-of-Thought connect language, image, and video models in film pipelines (Zhu et al., 2023; He et al., 2023; Zheng et al., 2025b), while AniMaker searches over candidate clips (Shi et al., 2025). ViMax, the closest system, combines narrative retrieval, a camera-dependency graph that passes parent frames to child cameras, and best-of-k keyframe selection (Huang et al., 2026a). ReCA allocates context across planning and frozen-generator calls (Liu et al., 2026). CANVAS specifies four decisions before rendering each shot: which pixels it inherits (at most one committed frame; Equation 2), which story state it receives as symbols (H_i), which object states must hold (Equation 3), and how many denoising steps it receives (Equation 4).

Feedback, inference compute, and evaluation. VideoAgent and NEWTON use feedback and verifiers to refine or replan generation (Soni et al., 2024; Feng et al., 2026), building on reflection-based agents (Yao et al., 2022; Shinn et al., 2023). Faster samplers and test-time scaling alter the cost of individual calls (Song et al., 2020; Lu et al., 2022; Liu et al., 2025). CANVAS allocates a fixed denoising budget across shots and uses typed verification: shots that fail a gate remain labeled degraded. We evaluate with VBench-style measures, ViCLIP, ArcFace, DINOv2, optical flow, and physics benchmarks (Huang et al., 2024; He et al., 2024b; Wang et al., 2023d; Deng et al., 2018; Oquab et al., 2023; Teed & Deng, 2020; Bansal et al., 2024). Following ViMax-Bench (Huang et al., 2026a), we pair comparisons by story and add per-shot face identity, within-scene seams, and VBench quality. Broader suites such as EvalCrafter and VBench-2.0 primarily evaluate individual clips (Liu et al., 2024; Zheng et al., 2025a).

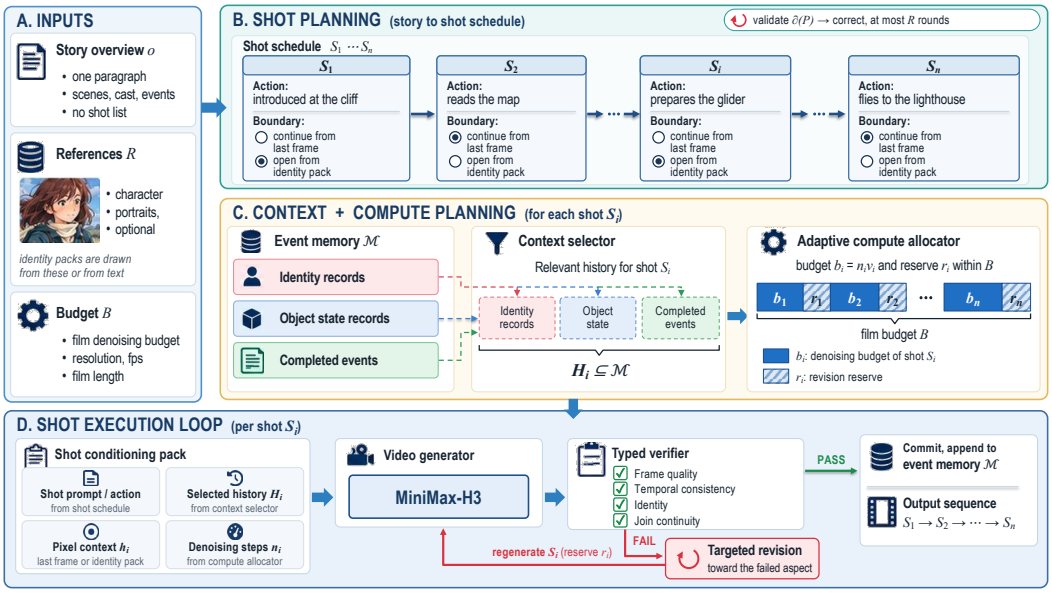


Figure 3: **CANVAS architecture.** (A) Inputs: overview o, optional references $\mathcal{R}$, and budget B. (B) Shot planning compiles a validated schedule with boundary labels (Equation 1). (C) Per shot, the planner selects history $H_i \subseteq \mathcal{M}$ and allocates steps n_i (cost $b_i = n_i v_i$) and revision reserve r_i. (D) MiniMax-H3 renders the shot from its visual input h_i, H_i, and n_i; a take that fails a gate is revised toward the failed check. On the identity-pack route, identity and join continuity are advisory; only frame quality and temporal consistency gate.

3 METHOD

3.1 PROBLEM SETUP

A shot that inherits the wrong context or budget is cheap to fix in text and costly to find in video, so CANVAS fixes both before generation. It takes a story overview o, optional character references $\mathcal{R}$, and a film-level denoising budget B, and produces a plan $P = (S_1, \ldots, S_N)$ and one committed take $x_i^\star$ per shot, logging every take's inputs, settings, and verifier reports. Each shot S_i records its scene, setting, cast c_i, duration, action, and object contract $\mathcal{O}_i$, and the plan fixes four decisions per shot: a *boundary label* $\sigma_i \in \{0, 1\}$, where $\sigma_i = 1$ continues from the last frame of $x_{i-1}^\star$ and $\sigma_i = 0$ opens from identity references ($\sigma_1 = 0$); a *history* $H_i \subseteq \mathcal{M}_{<i}$, the records of earlier shots that the shot is written against; a number of *denoising steps* n_i; and a *revision reserve* r_i, the number of extra attempts allowed if its take fails verification.

3.2 COMPILE A STORY INTO AN EXECUTABLE PLAN

A long shot list written in one pass drifts: costumes change mid-scene and shot counts stop matching the request. Compiling scene by scene and validating the result catches these errors while they are still text edits. An outline step writes scenes $\mathcal{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_K)$ with fixed cast and setting records, an allocation step splits the N shots across scenes, and a realization step writes each scene's shots. A validator $\partial(P)$ lists the violated constraints, and a correction step repairs them for at most R rounds:

$$\begin{aligned} P^{(0)} &= \mathrm{Concat}_{k=1}^{K} \textsc{Realize}(\mathcal{S}_k, N_k), \\ (N_1, \ldots, N_K) &= \textsc{Alloc}(\mathcal{S}, N), \qquad \mathcal{S} = \textsc{Outline}(o; \mathcal{R}), \\ P^{(j+1)} &= \textsc{Correct}\big(P^{(j)}, \partial(P^{(j)})\big) \quad \text{while } \partial(P^{(j)}) \neq \varnothing \text{ and } j < R, \end{aligned} \tag{1}$$

where Concat concatenates the scene plans in order. A plan that is still invalid after R rounds is rejected before rendering.

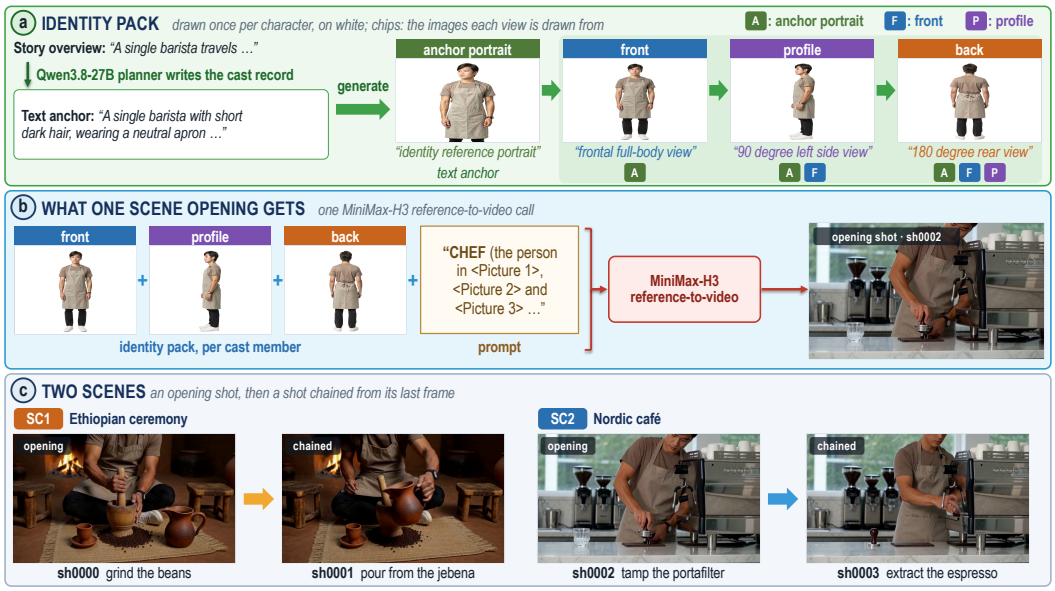


Figure 4: **Identity pack and scene openings.** (a) From the planner's text anchor, FLUX.2-klein draws an anchor portrait (A), then front, profile, and back views, each from its view words and the images marked below it. (b) One MiniMax-H3 reference-to-video call takes each cast member's three views and a prompt that binds people to pictures. (c) Two scenes of the evaluated barista film: an opening shot, then a shot chained from its last frame.

3.3 Route context by shot boundaries

Inside a continuous action, the previous last frame is the most direct continuity signal; across a cut, it drags the old room and pose into the new scene. Facts that must survive a cut, such as who holds what, need no pixels, so we route the two separately. The routing operator Γ gives each shot a pixel context h_i and a text history H_i:

$$\Gamma(S_i) = (h_i,\ H_i), \qquad h_i = \begin{cases} \mathrm{last}(x^{\star}_{i-1}), & \sigma_i = 1, \\ \mathcal{V}_{c_i}, & \sigma_i = 0, \end{cases} \qquad H_i = \textsc{Select}(\mathcal{M}_{<i}, S_i), \tag{2}$$

where $\mathcal{V}_{c_i}$ holds the front, profile, and back views of each cast member (Section 3.4; empty for a shot without characters) and $\mathcal{M}_{<i}$ holds the records of earlier shots. No pixels cross a scene cut, so a shot inherits at most one earlier frame regardless of film length. The planner chooses H_i while writing the plan and uses the selected records to specify the shot's objects, starting states, and action.

3.4 Identity packs and scene openings

Since no frame crosses a cut, each scene opening must re-establish its characters. A single frontal portrait leaves profile and back views unconstrained. Figure 4 shows the $\sigma_i = 0$ branch of Equation 2. Each character's identity pack (front, profile, and back views) is drawn once by FLUX.2-klein (Black Forest Labs, 2026) from the planner's text anchor. A scene opening is one MiniMax-H3 reference-to-video call on the packs of its cast, with a prompt that binds each person to their pictures. A cast-free opening uses the prompt alone.

3.5 Specify finite object-action contracts

Many continuity errors are object errors, such as a cup appearing in an empty hand. They are hard to detect in pixels but easy to state as positions, so a contract makes each hand-off checkable before rendering. Let $\mathcal{U}$ be the objects and Λ the locations (supports, containers, and hands). A state $z : \mathcal{U} \to \Lambda$ places every object, and an action $a = (u, \ell, \ell')$ moves object u from ℓ to ℓ' within a

time window. An action changes the state only if it is feasible:

$$\delta(z,a)=\begin{cases} z[u\mapsto \ell'], & z(u)=\ell,\ \ell\neq\ell',\ \mathrm{cap}(\ell')>\left|z^{-1}(\ell')\right|,\\ \bot, & \text{otherwise,}\end{cases} \tag{3}$$

where $\mathrm{cap}(\ell')$ is a location's capacity (e.g., one object per hand). A contract $\mathcal{O}_i = (z_i^0, a_1, \ldots, a_M, z_i^T)$ is valid if applying the actions in order never reaches $\bot$ and ends in z_i^T, object containment has no cycles, and a continuation starts where its predecessor ended ($\sigma_i = 1 \Rightarrow z_i^0 = z_{i-1}^T$). Each generation call receives only the actions that overlap its time window.

3.6 Allocate compute before generation

A uniform schedule gives a static establishing shot the same steps as a two-person hand-off. Because the plan already describes each shot, we set its steps from that description. A lexicon classifier maps the text of S_i to features $\phi(S_i)$ such as the number of subjects, interaction, motion, camera, and lighting (weights in Appendix D). A clipped weighted sum gives a complexity score κ_i, which four thresholds ϑ_j cut into five tiers:

$$\kappa_i=\mathrm{clip}_{[0,1]}\Big(\tfrac{1}{Z}\textstyle\sum_f w_f\,\phi_f(S_i)\Big),\qquad t_i=1+\textstyle\sum_{j=1}^{4}\mathbb{1}\big[\kappa_i\geq\vartheta_j\big],\qquad n_i=\mathcal{N}(t_i). \tag{4}$$

The step table $\mathcal{N}$ increases with the tier, by default $\mathcal{N}=(13,16,19,22,25)$. The revision reserve r_i can also depend on the tier, but every identity-pack run uses $r_i=2$. Shot i costs $b_i=n_i v_i$, where the volume v_i is its width $\times$ height $\times$ frame count relative to a reference shape. The film's planned cost is $\widehat{C}(P)=\sum_i b_i$. In the reported runs, B is the cost of a uniform 25-step pass over the same shots, and Figure 5a reports denoising spent relative to it.

3.7 Verify, revise, and commit a take

Some takes fail in ways a cheap check detects, such as frozen motion or corrupted frames. Instead of retrying blindly, the reviser uses the failed check to change the prompt. The verifier checks each take x on four aspects α: frame quality, temporal consistency, identity, and join continuity. For each aspect it computes a normalized score $q_\alpha(x)$ and a verdict $p_\alpha(x)\in\{\textsc{pass},\textsc{fail},\textsc{n/a}\}$ against a threshold τ_α. Only the active gates $\mathcal{A}_{\mathrm{act}}$ (frame quality and temporal consistency in the identity-pack runs) decide acceptance, and N/A does not count as a pass:

$$g(x)=\textstyle\bigwedge_{\alpha\in\mathcal{A}_{\mathrm{act}}}\mathbb{1}\big[p_\alpha(x)=\textsc{pass}\big]. \tag{5}$$

If a take fails, the reviser rewrites the prompt toward the worst failed aspect $\alpha^\dagger = \arg\min_{\alpha:\,p_\alpha=\textsc{fail}} q_\alpha$, keeps the event and character anchors, and renders again with the same context and steps. With prompt $\pi_i^{(k)}$ at attempt k and generator G, shot i uses at most $1+r_i$ attempts:

$$x_i^{(k+1)}=G\big(\textsc{Revise}(\pi_i^{(k)},\alpha^\dagger),\,\Gamma(S_i),\,n_i\big),\qquad k\leq r_i. \tag{6}$$

The first passing attempt is committed as $x_i^\star$. If none passes, the best-ranked attempt is committed as a *failed shot*, which keeps its failed verdicts and is scored like any other shot. Each committed $x_i^\star$ is logged with its prompt, settings, and report before S_{i+1} starts, and its end state is added to $\mathcal{M}$.

4 Experiments

4.1 Setup

Benchmark and metrics. ViMax-Bench contains 12 medium and 23 long multi-shot stories (Huang et al., 2026a). Following it, we report ViCLIP (Wang et al., 2023d) similarity between shots within a scene (IC), across scenes (CC), and over all shot pairs (GC). Each story counts equally, overall scores weight the Medium and Long groups by size, and 95% intervals resample stories. For every method, we drop near-copy shot pairs (ViCLIP cosine above 0.95), which chaining produces. Canvas writes its own scenes (5.3 per film, against 2–4 in the benchmark), so its IC and CC cover different pairs from the baselines'.

Table 1: **Consistency on ViMax-Bench** (CC/IC/GC: cross-scene, intra-scene, global ViCLIP consistency, ↑). GC compares all shot pairs and is comparable across systems; CC and IC use method-specific scene partitions. *Top:* published (Huang et al., 2026a, v2, Table 1), with Gemini-3-Flash (Google DeepMind, 2025a), Veo 3.1 (Google DeepMind, 2025b), and Wan2.2 (Team Wan et al., 2025). *Bottom:* our runs with shared planner, image model, and generator.

			Medium			Long			Overall		
Method	LLM planner	Video	CC	IC	GC	CC	IC	GC	CC	IC	GC
Published results on ViMax-Bench											
Wan2.2	–	Wan2.2	0.439	0.446	0.442	0.419	0.421	0.419	0.426	0.429	0.427
Veo 3.1	–	Veo 3.1	0.554	0.564	0.558	0.469	0.478	0.470	0.498	0.507	0.501
IC-LoRA	–	Wan2.2	0.540	0.571	0.554	0.502	0.533	0.509	0.515	0.546	0.524
StoryDiffusion	–	Wan2.2	0.561	0.580	0.569	0.501	0.520	0.505	0.521	0.540	0.527
HoloCine	–	HoloCine	0.537	0.589	0.560	0.494	0.557	0.508	0.510	0.569	0.527
MovieAgent	native	Veo 3.1	0.476	0.534	0.502	0.441	0.509	0.456	0.453	0.517	0.472
AniMaker	native	Veo 3.1	0.552	0.601	0.574	0.477	0.522	0.488	0.503	0.549	0.517
ViMax	Gemini-3-Flash	Wan2.2	0.564	0.590	0.575	0.498	0.547	0.509	0.521	0.561	0.532
ViMax (3.1-Flash img.)	Gemini-3-Flash	Veo 3.1	0.568	0.591	0.579	0.509	0.540	0.516	0.529	0.558	0.537
ViMax	Gemini-3-Flash	Veo 3.1	0.589	0.605	0.596	0.530	0.563	0.537	0.550	0.577	0.557
Our runs, shared planner and generator (35 stories)											
MovieAgent	Qwen3.8-27B	H3	0.473	0.524	0.476	0.420	0.475	0.440	0.438	0.492	0.452
ViMax	Qwen3.8-27B	H3	0.548	0.580	0.571	0.484	0.548	0.506	0.506	0.559	0.528
CANVAS (8–13 steps)	Qwen3.8-27B	H3	0.684	0.775	**0.701**	0.641	0.765	**0.666**	0.656	0.768	**0.678**
CANVAS (13–25 steps)	Qwen3.8-27B	H3	0.704	0.781	**0.716**	0.662	0.769	**0.683**	0.676	0.773	**0.694**
CANVAS, 25 steps every shot	Qwen3.8-27B	H3	0.708	0.782	**0.718**	0.657	0.768	**0.680**	0.675	0.773	**0.693**

Table 2: **Standard consistency and quality metrics** (35 stories; CANVAS 13–25 steps per shot). Mean and paired difference by story [95% CI]; blue: CANVAS higher, CI excluding 0. CANVAS, 8–13: adaptive tiers capped at 13 steps (values only).

Metric	CANVAS	CANVAS, 8–13	ViMax	MovieAg.	Δ ViMax	Δ MovieAg.
Consistency						
DINOv2 GC	**0.592**	0.567	0.297	0.243	**+0.295 [0.261, 0.330]**	**+0.349 [0.306, 0.393]**
Face identity (ArcFace)	**0.398**	0.406	0.187	0.225	**+0.210 [0.176, 0.245]**	**+0.176 [0.136, 0.216]**
Visual quality (VBench, per shot)						
Aesthetic quality (AQ)	0.583	0.578	0.572	0.561	+0.011 [−0.008, 0.030]	+0.022 [−0.001, 0.045]
Imaging quality (IQ)	0.689	0.629	0.686	0.673	+0.003 [−0.002, 0.020]	+0.016 [−0.0001, 0.028]
Subject consistency (SC)	**0.960**	0.959	0.945	0.920	**+0.015 [0.008, 0.022]**	**+0.040 [0.030, 0.049]**
Temporal flickering (TF)	**0.992**	0.994	0.990	0.981	**+0.002 [0.000, 0.004]**	**+0.011 [0.008, 0.013]**

Implementation details. All methods share Qwen3.8-27B (Qwen Team, 2026) as the language model, FLUX.2-klein-9B as the image model, and MiniMax-H3 (MiniMax, 2026) (H3) as the video generator, at 832×480 and 24 fps. (MovieAgent's keyframes come from its own StoryDiffusion module). Baselines run from their public code (ViMax with two disclosed changes) with 25 denoising steps per shot; CANVAS uses at most 25. CANVAS renders 437 single-call shots of 243 frames (10.125 s), with tiers of $\mathcal{N} = (13, 16, 19, 22, 25)$ steps by default and up to three attempts per shot. The main evaluation uses all 35 stories; the ablations and planner comparison use ten.

What the comparison controls. The frameworks share planner, generators, resolution, frame rate, shot length (243 frames, 10.125 s), maximum steps, and shots per story, but differ in scene partition, chaining, which raises within-scene similarity by construction, and acceptance: 17 of the 437 delivered CANVAS shots fail their own gates and are still scored. GC ignores scene labels, so scene partition cannot move it.

4.2 MAIN RESULTS

Consistency. Global consistency (GC), which covers all shot pairs irrespective of scene labels, is higher for each CANVAS setting than for the shared-backend baselines (Table 1; styles: Figure 6). The default CANVAS exceeds ViMax by +0.166 GC and MovieAgent by +0.242 GC. With our shared planner, image model, and generator, both baselines score below their published GC values (ViMax 0.528 vs. 0.532–0.557, MovieAgent 0.452 vs. 0.472).

Other evaluators and human study. DINOv2 (Oquab et al., 2023) and ArcFace face identity (Deng et al., 2018) agree with ViCLIP: CANVAS leads both baselines with intervals that exclude zero (Table 2). VBench subject consistency and temporal flickering are also higher, and aesthetic and imaging quality stay close to ViMax. Consistency is not bought with static shots: VBench dynamic degree is 0.40 against 0.34 for ViMax (MovieAgent 0.58), and per-shot alignment to the

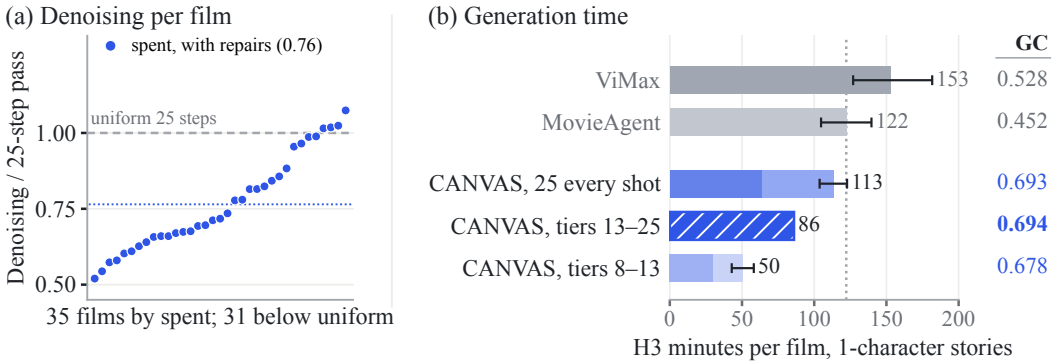


Figure 5: **Generation cost.** (a) Default CANVAS denoising spent per film, including repairs, relative to a uniform 25-step pass. (b) H3 minutes per film (one 48 GB RTX 6000 Ada, repairs included). Darker segments: openings; whiskers: 95% intervals across stories. Right: GC from Table 1.

Table 3: **Identity packs for the baselines** (10 films). Baselines keep their own plans; their scene openings use CANVAS's identity packs through the same H3 reference-to-video call.

Method	Scene openings	GC ↑	IC ↑	AQ ↑	IQ ↑
MovieAgent	own keyframes	0.450	0.482	0.539	0.656
MovieAgent	CANVAS identity packs	0.482	0.490	0.548	0.672
ViMax	own keyframes	0.527	0.548	0.544	0.638
ViMax	CANVAS identity packs	0.563	0.554	0.554	0.674
CANVAS	identity packs	0.692	0.777	0.560	0.682

Table 4: **Removing one component at a time** (R4). GC and IC exclude near-duplicate pairs; IC is undefined when no pair remains (7–10 films per arm). Failed: failed shots out of 88; time: relative H3 minutes for film generation; –: not reported.

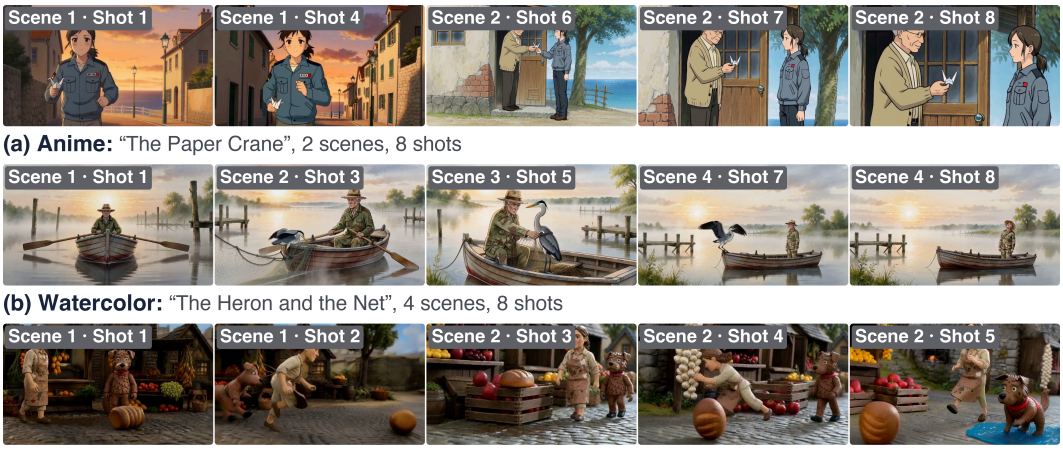
Variant	GC ↑	IC ↑	Failed ↓	AQ ↑	IQ ↑	Time ↓
w/o chaining (every shot opens)	0.680	0.741	6	0.556	0.680	1.55×
w/o identity references	0.659	0.765	5	0.546	0.647	0.64×
w/o object contracts (prose)	0.683	0.740	7	0.551	0.672	1.01×
uniform steps	0.688	0.781	5	0.563	0.687	1.32×
w/o typed revision (one take)	0.645	0.742	11	0.550	0.671	0.81×
untargeted resampling (three takes)	0.647	0.743	9	0.558	0.678	1.00×
CANVAS (all components)	0.692	0.777	5	0.560	0.682	1.00×

benchmark script is on par (Table 16). Three seeds on the ablation subset give GC 0.692, 0.702, and 0.697 (per-story SD 0.018). In a blinded study with 30 raters, 24 stories, and 240 pairwise judgments per baseline and criterion, raters prefer CANVAS over HoloCine, ViMax, and MovieAgent on every criterion (65–75% of judgments; Figure 7a).

Where the gap comes from. The GC gain over ViMax is larger on Long than Medium stories (+0.177 vs. +0.145). Across scenes, where no frame is carried over, CANVAS's CC is 0.676 against 0.506 for ViMax and 0.438 for MovieAgent. Giving the baselines CANVAS's identity packs (Table 3) raises GC by 0.036 for ViMax and 0.032 for MovieAgent, with slightly higher AQ and IQ and IC that changes by less than 0.01.

4.3 WHERE THE COMPUTE GOES

CANVAS's default schedule spends 0.76 times the denoising of a uniform 25-step pass across the 35-story benchmark, while GC is similar (0.694 vs. 0.693; Table 1; +0.0013 [−0.0047, 0.0069]). On the 13 single-character stories, H3 generation takes 86 minutes per film with the default schedule, compared with 153 for ViMax and 122 for MovieAgent; using 25 steps on every shot takes 113 minutes (Figures 1 and 5). The 8–13-step setting takes 50 minutes at GC 0.678 and IQ 0.629 (default: 0.694 and 0.689).



(a) Anime: "The Paper Crane", 2 scenes, 8 shots

(b) Watercolor: "The Heron and the Net", 4 scenes, 8 shots

(c) Claymation: "The Runaway Loaf", 3 scenes, 8 shots

Figure 6: **CANVAS across visual styles** at the default setting, one film per row: five shots in story order, one frame each. Stories are written for the figure; films and shots are selected.

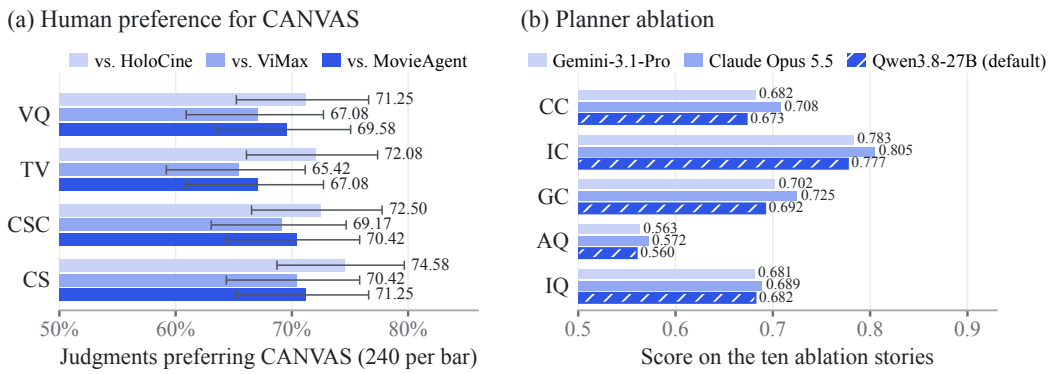


Figure 7: **Human evaluation and planner ablation.** (a) Share of 240 judgments per bar preferring CANVAS (VQ: visual quality; TV: text–video alignment; CSC: cross-scene and CS: character consistency); whiskers: Wilson 95% intervals. (b) Three planners on the ten ablation stories; Gemini-3.1-Pro and Claude Opus 5.5 lack per-story records.

4.4 CONTROLLED ABLATIONS

Each single-factor arm changes one component on the ten ablation stories (Table 4). Removing identity references changes GC from 0.692 to 0.659 and face match to the seed portraits from 0.152 to 0.014. Chaining makes the within-scene handoff smoother (48.6 without chaining, 3.8 with it) and raises IC from 0.741 to 0.777; GC is 0.676 without chaining and 0.692 with it. Replacing contracts with prose gives GC 0.683, close to 0.692 with contracts. One take leaves 11 failed shots, typed revision 5, and three untargeted takes 9 (GC 0.645, 0.692, 0.647). Other planners give GC 0.702 (Gemini-3.1-Pro) and 0.725 (Claude Opus 5.5) against 0.692 (Figure 7b).

5 CONCLUSION AND LIMITATIONS

We presented CANVAS, a training-free framework for routing visual and symbolic context across shot boundaries, with adaptive denoising as a separate efficiency feature. On ViMax-Bench, it raises global consistency from 0.528 for ViMax to 0.694 with the planner and generators held fixed. The tested ablations do not explain the full global-consistency gap: identity packs improve reference matching and chaining improves within-scene handoffs. Shot difficulty comes from a hand-set lexicon score, and the step tiers and verifier gates stay fixed within a run. Each shot's history is chosen at planning time, so the plan does not adapt to what earlier takes actually rendered. Identity-pack openings also require a generator that accepts reference images. Future work should isolate which planning choices account for the remaining gain and replan as takes are committed.

ETHICS STATEMENT

Reference-conditioned generation requires licensed inputs and authorization for identifiable likenesses. Generated work should be labeled; releases should exclude private references, credentials, and identifying paths. Human evaluation requires consent and privacy. All experimental characters are synthetic portraits generated from text, without real-person references. This framework is intended for authorized creative and research use, not deceptive impersonation.

AI USE STATEMENT

Generative AI assisted manuscript editing, implementation review, analysis code, and source checks. The water-cycle and character illustrations in Figures 2 and 3 came from AI-generated draft diagrams; labels and layout were typeset separately. Other pictures are outputs of evaluated systems. The authors reviewed the work and are responsible for its content.

REPRODUCIBILITY STATEMENT

Appendices B–C specify model versions, baseline changes, hardware, and execution settings; Appendices E and I describe ablations and evaluation. Reproduction requires frozen configurations, evaluated media, and execution records.

REFERENCES

Xiangyu Bai, He Liang, Bishoy Galoaa, Utsav Nandi, Shayda Moezzi, Yuhang He, and Sarah Ostadabbas. MoReGen: Multi-Agent Motion-Reasoning Engine for Code-based Text-to-Video Synthesis, 2025. URL https://arxiv.org/abs/2512.04221.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating Physical Commonsense for Video Generation, 2024. URL https://arxiv.org/abs/2406.03520.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation, 2025. URL https://arxiv.org/abs/2503.06800.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A Space-Time Diffusion Model for Video Generation, 2024. URL https://arxiv.org/abs/2401.12945.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence, January 2026. Model card: https://huggingface.co/black-forest-labs/FLUX.2-klein-9B.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets, 2023a. URL https://arxiv.org/abs/2311.15127.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models, 2023b. URL https://arxiv.org/abs/2304.08818.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation, 2023a. URL https://arxiv.org/abs/2310.19512.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models, 2024. URL https://arxiv.org/abs/2401.09047.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction, 2023b. URL https://arxiv.org/abs/2310.20700.

Nicolae Cudlenco, Mihai Masala, and Marius Leordeanu. Authoring for Living Worlds: Tool-Constrained LLM Agents for Executable Multi-Actor Scenarios, 2026. URL https://arxiv.org/abs/2604.10383.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition, 2018. URL https://arxiv.org/abs/1801.07698.

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Kaelbling, Andy Zeng, and Jonathan Tompson. Video Language Planning, 2023. URL https://arxiv.org/abs/2310.10625.

Yuxiang Feng, Juncheng Wang, Chao Xu, Yijie Qian, Huihan Wang, Wenlong Hou, Yang Liu, Baigui Sun, Yong Liu, and Shujun Wang. NEWTON: Agentic Planning for Physically Grounded Video Generation, 2026. URL https://arxiv.org/abs/2605.18396.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer, 2022. URL https://arxiv.org/abs/2204.03638.

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning, 2023. URL https://arxiv.org/abs/2311.10709.

Google DeepMind. Gemini 3 Flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf, December 2025a.

Google DeepMind. Veo 3.1. https://deepmind.google/models/veo/, 2025b. Model card: https://storage.googleapis.com/deepmind-media/Model-Cards/Veo-3-Model-Card.pdf; technical report: https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

Xu Guo, Zhengxuan Wei, Xinghui Li, Hanzhuo Huang, Xinyu Liu, Xiangyang Luo, Min Wei, Yiran Zhu, Qiulin Wang, Yulong Xu, Xintao Wang, Pengfei Wan, Qi Fan, and Xiangwang Hou. ContextMaster: Interactive Multi-Shot Video Creation via Fixed-Budget Sparse Context Routing, 2026. URL https://arxiv.org/abs/2608.04956.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning, 2023. URL https://arxiv.org/abs/2307.04725.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic Video Generation with Diffusion Models, 2023. URL https://arxiv.org/abs/2312.06662.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime Video Latent Diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible Diffusion Modeling of Long Videos, 2022. URL https://arxiv.org/abs/2205.11495.

Liu He, Yizhi Song, Hejun Huang, Pinxin Liu, Yunlong Tang, Daniel Aliaga, and Xin Zhou. Kubrick: Multimodal Agent Collaborations for Synthetic Video Generation, 2024a. URL https://arxiv.org/abs/2408.10453.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation, 2024b. URL https://arxiv.org/abs/2406.15252.

Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, and Qifeng Chen. Animate-A-Story: Storytelling with Retrieval-Augmented Video Generation, 2023. URL https://arxiv.org/abs/2307.06940.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2568–2577, 2025. doi: 10.1109/CVPR52734.2025.00245.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning, 2021. URL https://arxiv.org/abs/2104.08718.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, 2020. URL https://arxiv.org/abs/2006.11239.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High Definition Video Generation with Diffusion Models, 2022a. URL https://arxiv.org/abs/2210.02303.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video Diffusion Models, 2022b. URL https://arxiv.org/abs/2204.03458.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, 2022. URL https://arxiv.org/abs/2203.15556.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers, 2022. URL https://arxiv.org/abs/2205.15868.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. FilMaster: Bridging Cinematic Principles and Generative AI for Automated Film Generation, 2025. URL https://arxiv.org/abs/2506.18899.

Lingxuan Huang, Sizhe He, Hengji Zhou, Liqiang Nie, Lianghao Xia, and Chao Huang. ViMax: Agentic Video Generation, 2026a. URL https://arxiv.org/abs/2606.07649.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner Monologue: Embodied Reasoning through Planning with Language Models, 2022. URL https://arxiv.org/abs/2207.05608.

Yuyang Huang, Yabo Chen, Wenrui Dai, Ziyang Zheng, Haibin Huang, Chi Zhang, Junni Zou, Hongkai Xiong, and Xuelong Li. CineWeaver: Training-Free Reference-Controllable Multi-Shot Long Video Generation for Cinematic Storytelling, 2026b. URL https://arxiv.org/abs/2607.26529.

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. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024.

Kaixun Jiang, Yuzheng Wang, Junjie Zhou, Pandeng Li, Zhihang Liu, Chen-Wei Xie, Zhaoyu Chen, Yun Zheng, and Wenqiang Zhang. GenAgent: Scaling Text-to-Image Generation via Agentic Multimodal Reasoning, 2026. URL https://arxiv.org/abs/2601.18543.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How Far is Video Generation from World Model: A Physical Law Perspective, 2024. URL https://arxiv.org/abs/2411.02385.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. VideoPoet: A Large Language Model for Zero-Shot Video Generation, 2023. URL https://arxiv.org/abs/2312.14125.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A Systematic Framework For Large Video Generative Models, 2024. URL https://arxiv.org/abs/2412.03603.

Yixuan Lai, Tianjia Shao, Weijia Dou, Siyu Zhu, and Jingdong Wang. GroundShot: Visually Consistent Multi-Shot Long Video Generation via Entity-Grounded Shot Scheduling, 2026. URL https://arxiv.org/abs/2606.20799.

Chunhan Li, Qifeng Wu, Jia-Hui Pan, Ka-Hei Hui, Jingyu Hu, Yuming Jiang, Bin Sheng, Xihui Liu, Wenjuan Gong, and Zhengzhe Liu. coDrawAgents: A Multi-Agent Dialogue Framework for Compositional Image Generation, 2026. URL https://arxiv.org/abs/2603.12829.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. In-the-Flow Agentic System Optimization for Effective Planning and Tool Use, 2025. URL https://arxiv.org/abs/2510.05592.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent Multi-scene Video Generation via LLM-Guided Planning, 2023a. URL https://arxiv.org/abs/2309.15091.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, and Ming-Hsuan Yang. Text-Driven Image Editing via Learnable Regions, 2023b. URL https://arxiv.org/abs/2311.16432.

Yuanze Lin, Yunsheng Li, Dongdong Chen, Weijian Xu, Ronald Clark, and Philip H. S. Torr. Olympus: A Universal Task Router for Computer Vision Tasks, 2024. URL https://arxiv.org/abs/2412.09612.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Ronald Clark, and Ming-Hsuan Yang. IllumiCraft: Unified Geometry and Illumination Diffusion for Controllable Video Generation, 2025. URL https://arxiv.org/abs/2506.03150.

Akide Liu, Jinbo Xing, Chaojie Mao, Ye Li, Zeyu Zhang, Yefei He, Weijie Wang, Zihan Wang, Yu Liu, Gholamreza Haffari, and Bohan Zhuang. ReCA: Multi-Shot Long Video Extrapolation via Recursive Context Allocation, 2026. URL https://arxiv.org/abs/2605.26525.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-Time Scaling for Video Generation, 2025. URL https://arxiv.org/abs/2503.18942.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and Evaluating Large Video Generation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, 2024.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. FETV: A Benchmark for Fine-Grained Evaluation of Open-Domain Text-to-Video Generation, 2023. URL https://arxiv.org/abs/2311.01813.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps, 2022. URL https://arxiv.org/abs/2206.00927.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. FreeLong: Training-Free Long Video Generation with SpectralBlend Temporal Attention, 2024. URL https://arxiv.org/abs/2407.19918.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent Diffusion Transformer for Video Generation, 2024. URL https://arxiv.org/abs/2401.03048.

Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, and Sergey Tulyakov. Snap Video: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis, 2024. URL https://arxiv.org/abs/2402.14797.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards World Simulator: Crafting Physical Commonsense-Based Benchmark for Video Generation, 2024. URL https://arxiv.org/abs/2410.05363.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, Yujun Shen, and Huamin Qu. HoloCine: Holistic Generation of Cinematic Multi-Shot Long Video Narratives, 2025. URL https://arxiv.org/abs/2510.20822.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. https://www.minimax.io/blog/minimax-h3, July 2026. Model weights: https://huggingface.co/MiniMaxAI/MiniMax-H3.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2025. URL https://arxiv.org/abs/2501.09038.

OpenAI. GPT-4 Technical Report, 2023. URL https://arxiv.org/abs/2303.08774.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. https://qwen.ai/blog?id=qwen3.8, August 2026. Citation requested by the Qwen3.8-27B model card, https://huggingface.co/Qwen/Qwen3.8-27B.

Akshat Rakheja, Aarsh Ashdhir, Aryan Bhattacharjee, and Vanshika Sharma. World Consistency Score: A Unified Metric for Video Generation Quality, 2025. URL https://arxiv.org/abs/2508.00144.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, 2021. URL https://arxiv.org/abs/2112.10752.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, 2022. URL https://arxiv.org/abs/2208.12242.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools, 2023. URL https://arxiv.org/abs/2302.04761.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. AniMaker: Multi-Agent Animated Storytelling with MCTS-Driven Clip Generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025. doi: 10.1145/3757377.3764009.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning, 2023. URL https://arxiv.org/abs/2303.11366.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data, 2022. URL https://arxiv.org/abs/2209.14792.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models, 2020. URL https://arxiv.org/abs/2010.02502.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models, 2023. URL https://arxiv.org/abs/2303.01469.

Achint Soni, Sreyas Venkataraman, Abhranil Chandra, Sebastian Fischmeister, Percy Liang, Bo Dai, and Sherry Yang. VideoAgent: Self-Improving Video Generation, 2024. URL https://arxiv.org/abs/2410.10076.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models, 2025. URL https://arxiv.org/abs/2503.20314.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow, 2020. URL https://arxiv.org/abs/2003.12039.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-Free Consistent Text-to-Image Generation, 2024. URL https://arxiv.org/abs/2402.03286.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Haitao Mi, and Dong Yu. Toward Self-Improvement of LLMs via Imagination, Searching, and Criticizing, 2024. URL https://arxiv.org/abs/2404.12253.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. StableAnimator: High-Quality Identity-Preserving Human Image Animation, 2024. URL https://arxiv.org/abs/2411.17697.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable Length Video Generation From Open Domain Textual Description, 2022. URL https://arxiv.org/abs/2210.02399.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. WISA: World Simulator Assistant for Physics-Aware Text-to-Video Generation, 2025. URL https://arxiv.org/abs/2503.08153.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope Text-to-Video Technical Report, 2023a. URL https://arxiv.org/abs/2308.06571.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul Debevec, and Ning Yu. MAViS: A Multi-Agent Framework for Long-Sequence Video Storytelling. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2273–2295, 2026. doi: 10.18653/v1/2026.eacl-long.101.

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, daiyuan li, Yu Hu, and Mingkui Tan. Generating Long-form Story Using Dynamic Hierarchical Outlining with Memory-Enhancement, 2024. URL https://arxiv.org/abs/2412.13575.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional Video Synthesis with Motion Controllability, 2023b. URL https://arxiv.org/abs/2306.02018.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models, 2023c. URL https://arxiv.org/abs/2309.15103.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation, 2023d. URL https://arxiv.org/abs/2307.06942.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated Movie Generation via Multi-Agent CoT Planning, 2025. URL https://arxiv.org/abs/2503.07314.

Junfei Xiao, Feng Cheng, Lu Qi, Liangke Gui, Jiepeng Cen, Zhibei Ma, Alan Yuille, and Lu Jiang. VideoAuteur: Towards Long Narrative Video Generation, 2025. URL https://arxiv.org/abs/2501.06173.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors, 2023a. URL https://arxiv.org/abs/2310.12190.

Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. SimDA: Simple Diffusion Adapter for Efficient Video Generation, 2023b. URL https://arxiv.org/abs/2308.09710.

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. FilmAgent: A Multi-Agent Framework for End-to-End Film Automation in Virtual 3D Spaces, 2025. URL https://arxiv.org/abs/2501.12909.

Bangji Yang, Ruihan Guo, Jiajun Fan, Chaoran Cheng, and Ge Liu. M3: High-fidelity Text-to-Image Generation via Multi-Modal, Multi-Agent and Multi-Round Visual Reasoning, 2026. URL https://arxiv.org/abs/2602.06166.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer, 2024. URL https://arxiv.org/abs/2408.06072.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-And-Write: Towards Better Automatic Storytelling, 2018. URL https://arxiv.org/abs/1811.05701.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, 2022. URL https://arxiv.org/abs/2210.03629.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. DragNUWA: Fine-grained Control in Video Generation by Integrating Text, Image, and Trajectory, 2023a. URL https://arxiv.org/abs/2308.08089.

Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Gong Ming, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation, 2023b. URL https://arxiv.org/abs/2303.12346.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-Rewarding Language Models, 2024. URL https://arxiv.org/abs/2401.10020.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation, 2023a. URL https://arxiv.org/abs/2309.15818.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. StoryMem: Multi-shot Long Video Storytelling with Memory, 2025a. URL https://arxiv.org/abs/2512.19539.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models, 2023b. URL https://arxiv.org/abs/2302.05543.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. STAGE: Storyboard-Anchored Generation for Cinematic Multi-shot Narrative, 2025b. URL https://arxiv.org/abs/2512.12372.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. MovieDreamer: Hierarchical Generation for Coherent Long Visual Sequence, 2024. URL https://arxiv.org/abs/2407.16655.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness, 2025a. URL https://arxiv.org/abs/2503.21755.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023. URL https://arxiv.org/abs/2306.05685.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention, 2025b. URL https://arxiv.org/abs/2503.15138.

Jinsong Zhou, Yihua Du, Xinli Xu, Luozhou Wang, Zijie Zhuang, Yehang Zhang, Shuaibo Li, Xiaojun Hu, Bolan Su, and Ying-cong Chen. VideoMemory: Toward Consistent Video Generation via Memory Integration, 2026. URL https://arxiv.org/abs/2601.03655.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation, 2024. URL https://arxiv.org/abs/2405.01434.

Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. MovieFactory: Automatic Movie Creation from Text using Large Generative Models for Language and Images, 2023. URL https://arxiv.org/abs/2306.07257.

Appendix

Supplementary material for

A Glossary and run registry 20

B Implementation details 20

C Baselines and released materials 21

D Complexity score and budget 21

E Ablations 21

F Controls, quality, and repetition 22

G 13 against 25 denoising steps 24

H Human study and planner ablation 24

I Metric definitions 25

J Qualitative examples 26

K Execution procedure and an example contract 27

L Extended related work 27

A GLOSSARY AND RUN REGISTRY

Terms. *Opening*: the first shot of a scene. *Identity pack*: one character's front, profile, and back views on white; an *identity-pack opening* is rendered from the packs of its cast. *Continuation*: a shot that starts from the previous shot's last frame (chaining). *Failed shot*: a shot delivered after its last attempt still fails a gate; it is scored like any other shot. *Default*: difficulty tiers of 13–25 steps. *Near-duplicate filter*: CC, IC, and GC drop shot pairs with ViCLIP cosine above 0.95, for every system; tables marked *all pairs* keep them.

Route tags. Every caption starts with the route it reports. is the *identity-pack route*, the system of the main text: openings are rendered from identity packs by H3's reference-to-video workflow. is the *earlier route*: openings start from a FLUX-composed frame (Appendix B). The two routes share the planner, contracts, tiers, and verifier, but differ in opening images and whether identity is a gate. Results marked are an earlier route and are not direct evidence for the main system. Table 5 lists every run.

Table 5: **Run registry.** r: revision reserve per shot (one attempt plus r repairs); candidates: videos per attempt. All CANVAS runs use one seed per film unless noted.

	Run	Route	Steps per tier	r	Cand.	Stories	Used in
R1	Default		13, 16, 19, 22, 25	2	1	35	Tabs. 1, 2, 10, 11; Figs. 1, 5
R2	25 every shot		25	2	1	35	Tabs. 1, 13, 12
R3	Tiers 8–13		8, 8, 9, 10, 13	2	1	35	Tabs. 1, 2, 13
R4	Single-factor arms		as R1, one factor changed	2	1	10	Tabs. 4, 6
R5	Seeds 2 and 3		as R1	2	1	10	Tab. 7
R6	Planner ablation		as R1	2	1	10	Fig. 7b, Tab. 14
E1	Tiers 8–13		8, 8, 9, 10, 13	0–3 by tier	1 (2 extreme)	35	Tab. 13
E3	Ablation arms		13, 16, 19, 22, 25	2	1	10	Tab. 8
B1	ViMax, MovieAgent	–	25 (243 frames per shot)	–	–	35	Tabs. 1, 11
B2	ViMax, MovieAgent with CANVAS packs	–	25 (openings from R1 packs)	–	–	10	Tab. 3

R1–R3 differ only in the step policy. All three give every shot two repairs and one candidate, with no early stopping, on the same 35 plans, identity packs, and seeds. Results on a subset say so in the caption; the ablation stories are the same ten (R4–R6 and E3). R4 changes one factor at a time.

B IMPLEMENTATION DETAILS

Models. All systems share Qwen3.8-27B (bfloat16, vLLM, tensor parallelism 2, snapshot `1d4bf0f`, temperature 0.2), MiniMax-H3 (4-bit SDNQ weights, revision `c38d07f`), and FLUX.2-klein-9B (revision `92196c8`). H3 renders 832×480 at 24 fps; a CANVAS shot is one 243-frame call (10.125 s), guidance 1.0.

Identity packs and openings . FLUX.2-klein draws an anchor portrait from the planner's text anchor, then the front, profile, and back views (28 steps, guidance 4.0), each matted to white. An opening is one H3 reference-to-video call on the cast's views in order (at most nine pictures; back, then profile views of the last-billed characters are dropped first), with a prompt that binds each person to their pictures; a cast-free opening uses the prompt alone.

Openings . FLUX composes the opening frame from each character's seed portrait and face crops (at most three references, up to three draws until ArcFace similarity ≥ 0.35, plus a wardrobe check); H3 then animates it.

Verifier and reviser. Gates: frame quality and temporal consistency. Advisory checks: identity (ArcFace) and join continuity; on the identity check was a gate. The reviser applies typed, rule-based prompt rewrites and may change the seed, never the step count; when attempts run out it keeps the best-ranked take as a failed shot.

Seeds. Each shot's seed is a fixed function of the film, its authored seed, and the attempt index, so a rerun receives the same seeds; H3 is deterministic given its inputs.

Style films (Figures 1 and 6). Each film comes from a one-sentence story prompt written for the figure, planned by Qwen3.8-27B with eight requested shots and rendered at the default setting, with identity packs drawn in the film's style. These are hand-authored visual demonstrations, selected for style diversity rather than sampled from ViMax-Bench. Displayed shots are selected for illustration.

C BASELINES AND RELEASED MATERIALS

Baselines. ViMax runs from its public code (commit `05a4894`) with two changes: the camera-tree call is constrained to ViMax's own response schema with one identical retry (unmodified, about half of its camera trees failed to parse), and its prompts pass only the annotated scene and shot counts. MovieAgent runs from its public code (commit `d84041b`) with its client pointed at the shared planner, FLUX.2-klein portraits as its photo bank, and its StoryDiffusion keyframe module called directly, because the released code does not wire it in. Each shared-backend baseline shot is one 243-frame H3 image-to-video call (10.125 s); ViMax's 1600×896 keyframes are resized to 832×480.

Released materials. For every film: the authored plan, seeds, identity packs, effective configuration, per-shot prompts and conditioning inputs, verifier reports, and per-shot clips; the scoring code and per-story scores behind every table; and the code release, whose README documents hash matching and archive provenance. The 35 plans were composed once and are stored; every run renders from them.

D COMPLEXITY SCORE AND BUDGET

Features. Each feature adds a fixed weight to a raw sum, which is divided by 1.25 and clipped to $[0, 1]$. Subjects: 0, 0.06, 0.14, 0.20, 0.26 for 0–4 or more. Interaction: none 0, proximity 0.05, contact 0.13, manipulation 0.17, crowd 0.20. Motion: static 0, subtle 0.04, moderate 0.11, fast 0.19, chaotic 0.26. Camera: static 0, simple 0.06, complex 0.14. Lighting: static 0, gradual 0.05, dynamic 0.11. Background: empty 0, simple 0.02, detailed 0.06, crowd 0.13. Physics events 0.055 each (cap 0.18); visual effects 0.045 each (cap 0.14); face-critical 0.08; hand-critical 0.09; on-screen text 0.05; continuation 0.16; 0.035 per dependent later shot (cap 0.14); 0.012 per second beyond 4 s. Thresholds $\vartheta = (0.60, 0.65, 0.71, 0.79)$ cut the score into five tiers. The weights are hand-set priors about where video generators fail, not fitted parameters; the continuation weight was raised after inspecting a 20 s development film outside ViMax-Bench.

E ABLATIONS

Protocol . Each single-factor arm changes one factor of R1 on ten ViMax-Bench stories (41–52 s, 8–10 shots), fixed before any ablation render, with the same plans, packs, and seeds. Differences are paired by story with 95% bootstrap intervals over stories. Table 6 gives each component's effect on its target measure; Table 7 gives seed variation.

Table 6: **What each component buys**. Difference: CANVAS minus the variant, paired by story [95% CI] where available; blue bold: favors CANVAS, interval excluding 0. IC excludes near-duplicates; the paired IC contrast uses nine films with an eligible pair in both arms.

Variant	Target measure	CANVAS	Variant	Difference [95% CI]	GC (ours / variant)
No chaining	Within-scene handoff ↓	**3.8**	48.6	**−44.7 [−51.4, −38.2]**	0.692 / 0.676
	IC (intra-scene) ↑	0.777	0.741	−0.036 [−0.045, 0.005]	
No identity references	Face match to seed portraits ↑	**0.152**	0.014	**+0.138 [0.099, 0.174]**	0.692 / 0.659
No typed revision (one take)	Failed shots ↓	**5**	11	**−0.60 [−1.20, −0.10]**	0.692 / 0.645
Untargeted resampling, three takes	Failed shots ↓	5	9	−0.30 [−0.50, −0.03]	0.692 / 0.647
Uniform steps	Denoising spent ↓	**0.76×**	1.32×	**−0.56 [−0.71, −0.44]**	0.692 / 0.688

Chaining reduces within-scene handoff change from 48.6 without chaining to 3.8 with chaining (Table 6).

Table 7: **Seed variation** (R1 and R5).

	Seed 1 (R1)	Seed 2	Seed 3	Per-story SD, mean (max)
GC	0.692	0.702	0.697	0.018 (0.046)

Object-state fidelity . Qwen3.8-27B answers yes or no for each object-plan statement on single frames (40 transfer and 274 end-state statements). The share shown true is 0.743 with contracts and 0.721 with free prose. Contracts do not raise measured object-state fidelity. The full arm has 5 failed shots against 7 with prose on 88 shots, a two-shot difference whose paired story interval includes zero (-0.20 [-0.70, 0.20]); this is not evidence of fewer failures.

Execution and repairs. R1 made 512 video attempts for 437 shots. 40 shots failed their first attempt; 23 passed after revision and 17 were delivered as failed shots. First-attempt pass rates do not fall with tier (0.91, 0.88, 0.89, 0.88, 0.96 from trivial to extreme), so the tiers do not predict what fails. On E1 the reviser logged 47 repairs: 2 fixed the failed check, 44 changed nothing, and 1 made it worse; on E3 the identity gate failed all 27 of its checks because faces are small at 832×480, which is why identity is advisory on .

Table 8: **Single-factor ablations** (E3; earlier route, distinct from the main-paper Table 4; IC over all pairs, no near-duplicate filter). ΔGC is paired by story against CANVAS. Intervals are shown where supplied; per-story records for the revised aggregates are unavailable for verification or recomputation.

Arm	CC	IC	GC	ΔGC [95% CI]	GPU min
CANVAS (tiers 13–25, typed revision, contracts)	0.650	0.892	0.693	–	57.1
(i) Uniform steps	0.649	0.897	0.694	$+0.001$ [-0.005, 0.007]	74.2
(ii) Wide tiers 7/12/16/21/25	0.643	0.891	0.692	-0.001 [-0.002, 0.003]	41.1
(iii) Narrow tiers 19/21/22/24/25	0.657	0.901	0.700	$+0.007$ [-0.002, 0.017]	61.4
(iv) Wan2.2-TI2V-5B backbone	0.650	0.899	0.691	-0.003 [-0.035, 0.034]	33.4
(v) No chaining	0.620	0.821	0.668	-0.025 [-0.030, 0.020]	88.5
(vi) Single take	0.621	0.861	0.667	-0.026 [-0.032, 0.016]	43.2
(vii) Untargeted resample, 3 takes	0.612	0.863	0.666	-0.027 [-0.041, 0.015]	49.3
(viii) Free-prose object state	0.611	0.862	0.664	-0.029 [-0.043, 0.012]	48.8

On , uniform steps take 74.2 GPU minutes; no chaining reaches GC 0.668 at 88.5 GPU minutes; single take, three-take resampling, and free-prose object state reach GC 0.667, 0.666, and 0.664. Per-story records for revised aggregate rows are unavailable to verify or recompute their paired intervals.

Video-generator comparison. Table 9 gives exploratory alternate-backbone aggregates (VBench quality above, ViCLIP consistency below). These earlier-route rows use different opening images and identity gating, so they are not matched comparisons with the system. The MiniMax-H3 comparator is shown separately using main-paper R4 values (GC 0.692, AQ 0.560, IQ 0.682, CC 0.677, and IC 0.777; Tables 4, 3, and 14) and Table 2 quality values (SC 0.960, TF 0.992). Per-story records for the alternate backbones are unavailable, so paired comparisons cannot be computed.

Table 9: **Exploratory alternate-backbone aggregates** on the earlier route. The separate MiniMax-H3 comparator uses main-paper R4 consistency values and Table 2 SC/TF values; openings and identity gating differ, so these rows are not a matched comparison. Paired uncertainty is unavailable.

Generator	GC	AQ	IQ	SC	TF
MiniMax-H3, default (reference)	0.692	0.560	0.682	0.960	0.992
Wan2.2-TI2V-5B, 20 steps	0.688	0.548	0.717	0.937	0.974
LTX-2.3 distilled, 8 steps	0.620	0.554	0.710	0.922	0.980
Generator	**CC**	**IC**	**GC**		
MiniMax-H3, default (reference)	0.677	0.777	0.692		
Wan2.2-TI2V-5B, 20 steps	0.643	0.891	0.688		
LTX-2.3 distilled, 8 steps	0.601	0.739	0.620		

F CONTROLS, QUALITY, AND REPETITION

All shared-backend shots have the same nominal duration (243 frames at 24 fps). The table reports within-run CANVAS controls; the headline system comparisons are in Table 1.

Table 10: **Within-run consistency controls.** R1 spans 35 stories; †the ten ablation stories. GC is the story-equal ViCLIP mean. Each difference is variant minus the matching default over the same stories, with a paired story-bootstrap 95% interval where available; the corrected no-chaining aggregate has no recoverable paired interval.

Condition	Stories	GC	Change from R1 [95% CI]
R1, near-duplicates excluded (Table 1)	35	0.694	–
R1, all pairs retained	35	0.718	+0.024 [0.018, 0.029]
R1, failed shots excluded	35	0.696	+0.002 [−0.001, 0.005]
R1 on ablation subset†	10	0.692	–
No chaining†	10	0.676	−0.016 (CI unavailable)

Table 11: **Additional standard metrics** (35 stories; values match the main-paper Table 2). Mean and paired difference by story [95% CI].

Metric	CANVAS	ViMax	MovieAg.	Δ ViMax	Δ MovieAg.
DINOv2 GC	**0.592**	0.297	0.243	**+0.295 [0.261, 0.330]**	**+0.349 [0.306, 0.393]**
Face identity (ArcFace)	**0.398**	0.187	0.225	**+0.210 [0.176, 0.245]**	**+0.176 [0.136, 0.216]**
Aesthetic quality (AQ)	0.583	0.572	0.561	+0.011 [−0.008, 0.030]	+0.022 [−0.001, 0.045]
Imaging quality (IQ)	0.689	0.686	0.673	+0.003 [−0.002, 0.020]	+0.016 [−0.0001, 0.028]
Subject consistency (SC)	**0.960**	0.945	0.920	**+0.015 [0.008, 0.022]**	**+0.040 [0.030, 0.049]**
Temporal flickering (TF)	**0.992**	0.990	0.981	**+0.002 [0.000, 0.004]**	**+0.011 [0.008, 0.013]**

Table 12: **VBench quality against ViMax** (35 stories): paired difference per story [95% CI].

System − ViMax	AQ	IQ	SC	TF
CANVAS (R1)	+0.011 [−0.008, 0.030]	+0.003 [−0.002, 0.020]	+0.015 [0.008, 0.022]	+0.002 [0.000, 0.004]
CANVAS, 25 every shot (R2)	+0.001 [−0.020, 0.021]	+0.019 [0.004, 0.035]	+0.012 [0.004, 0.019]	+0.001 [−0.001, 0.003]
MovieAgent	−0.011 [−0.025, 0.002]	−0.013 [−0.026, 0.001]	−0.025 [−0.034, −0.016]	−0.009 [−0.010, −0.007]

G 13 AGAINST 25 DENOISING STEPS

Table 13: **13 against 25 steps.** Identity-pack route: R3 against R2 and R1 on 35 stories. Earlier opening route (‡): 35 stories at both settings. Near-duplicate filter applied. ΔGC and its 95% interval are paired by story; "25 higher" counts paired stories.

	13 steps			25 steps				
Comparison	CC	IC	GC	CC	IC	GC	ΔGC [95% CI]	25 higher
tiers 8–13 vs. every shot 25	0.656	0.768	0.678	0.675	0.773	0.693	+0.016 [0.003, 0.028]	24/35
tiers 8–13 vs. tiers 13–25	0.656	0.768	0.678	0.676	0.773	0.694	+0.017 [0.006, 0.027]	25/35
tiers 8–13 vs. every shot 25‡	0.647	0.769	0.676	0.649	0.766	0.679	+0.003 [−0.008, 0.015]	20/35
tiers 8–13 vs. tiers 13–25‡	0.647	0.769	0.676	0.646	0.763	0.673	−0.002 [−0.015, 0.010]	15/35

On , capping the tiers at 13 steps costs 0.017 GC. On the earlier opening route, the 35 paired films show +0.003 [−0.008, 0.015] GC against 25 steps on every shot and −0.002 [−0.015, 0.010] against tiers 13–25. Figure 1 compares R1 with the baselines at 25 steps.

H HUMAN STUDY AND PLANNER ABLATION

Human study design. Twenty-four ViMax-Bench stories (8 Medium, 16 Long) were drawn with a fixed seed (2027) before any rating. Each trial pairs CANVAS (R1) with one baseline on the same story: ViMax, MovieAgent (both B1), or HoloCine at its default settings. Thirty raters each judge all 24 stories, 8 against each baseline; rater r sees story i against baseline $(i + r) \bmod 3$, so every story–baseline cell is judged by 10 raters, giving 240 judgments per baseline and question. In each trial a rater reads the story overview, watches both films to the end with sound (the page unlocks the answers only then), and makes four two-alternative forced choices: visual quality (VQ), text–video alignment (TV), cross-scene consistency (CSC), and character consistency (CS). Films carry opaque names, are transcoded to identical H.264/AAC settings, and appear in a rater-seeded order and left–right position.

Quality control and consent. Each session adds an instructed-response trial ("choose B for every question"), whose failure excludes the rater, and one repeated trial with the sides swapped, which measures intra-rater agreement. Before starting, raters read an information page and tick a consent box: participation is voluntary, they may stop at any time and have their answers deleted before analysis, and we store only answers, trial times, and browser type under pseudonymous IDs.

Analysis. Figure 7a plots the reported shares. Each percentage corresponds to an integer count out of 240; Table 15 lists these counts with Wilson intervals and nominal two-sided binomial tests. These summaries treat judgments as independent; since raters and stories recur, we report them descriptively rather than as clustered uncertainty estimates.

Table 14: **Planner ablation** (R6): CANVAS planned by three language models, everything else as R1. The Qwen row is recomputed from R4 features; the other two rows are reported aggregates whose per-story records were unavailable for paired analysis.

Planner	CC	IC	GC	AQ	IQ
Qwen3.8-27B (R4 full arm)	0.677	0.777	0.692	0.560	0.682
Gemini-3.1-Pro	0.682	0.783	0.702	0.563	0.681
Claude Opus 5.5	0.708	0.805	0.725	0.572	0.689

Other stored measurements. Across the three R1 seeds on the ten ablation stories, matched shots have mean ViCLIP cosine distance 0.091 [0.078, 0.104] (story bootstrap); this measures variation across seeds, not motion within a film or narrative diversity. A GroundShot comparison is not available, and no single-factor arm isolates every individual property of the compiled plan.

Motion and script alignment . Table 16 tests whether consistency comes from static or off-script shots (R1 against B1, 35 stories). *Dynamic degree* is VBench's official measure: RAFT flow on frames sampled at 8 fps, and a shot counts as dynamic when enough frame pairs exceed a resolution-scaled threshold; *motion* is the median over frame pairs of the top-5% flow magnitude (pixels at 832×480). *Script alignment* is the ViCLIP cosine between shot k and the k-th shot of the benchmark script (its `video_prompt`); ViCLIP reads 32 tokens, so each script shot is embedded sentence by

Table 15: **Human preference, aggregate counts only.** Reported share for CANVAS is the integer count divided by 240. Wilson 95% intervals and exact two-sided binomial p values assume independent judgments; repeated raters and stories make them descriptive rather than clustered uncertainty.

Baseline	Criterion	Prefer CANVAS	Share	Wilson interval	Nominal p
HoloCine	VQ	171/240	71.25%	[65.2, 76.6]%	<0.001
HoloCine	TV	173/240	72.08%	[66.1, 77.4]%	<0.001
HoloCine	CSC	174/240	72.50%	[66.5, 77.8]%	<0.001
HoloCine	CS	179/240	74.58%	[68.7, 79.7]%	<0.001
ViMax	VQ	161/240	67.08%	[60.9, 72.7]%	<0.001
ViMax	TV	157/240	65.42%	[59.2, 71.1]%	<0.001
ViMax	CSC	166/240	69.17%	[63.1, 74.7]%	<0.001
ViMax	CS	169/240	70.42%	[64.4, 75.8]%	<0.001
MovieAgent	VQ	167/240	69.58%	[63.5, 75.1]%	<0.001
MovieAgent	TV	161/240	67.08%	[60.9, 72.7]%	<0.001
MovieAgent	CSC	169/240	70.42%	[64.4, 75.8]%	<0.001
MovieAgent	CS	171/240	71.25%	[65.2, 76.6]%	<0.001

sentence and averaged. Across films, GC is unrelated to motion within every system (Spearman ρ with dynamic degree: 0.07 for CANVAS, 0.00 for ViMax, 0.03 for MovieAgent).

Table 16: **Motion and script alignment** (35 stories; 437 shots per system). Mean [95% CI]; last rows: CANVAS minus the baseline, paired by story.

System	Dynamic degree ↑	Motion (px)	Script alignment ↑
CANVAS	0.401 [0.338, 0.473]	8.6 [6.2, 12.5]	0.123 [0.116, 0.129]
ViMax	0.340 [0.283, 0.396]	7.5 [6.0, 9.2]	0.117 [0.112, 0.123]
MovieAgent	0.584 [0.532, 0.637]	11.6 [10.4, 12.8]	0.117 [0.111, 0.122]
CANVAS − ViMax	+0.061 [−0.011, 0.144]	+1.1 [−1.4, 4.6]	+0.005 [0.001, 0.010]
CANVAS − MovieAgent	−0.184 [−0.261, −0.101]	−2.9 [−5.6, 0.8]	+0.006 [−0.000, 0.012]

I METRIC DEFINITIONS

ViCLIP consistency. For story j and method m, let s_u be the scene of delivered shot u and $\cos(x_u, x_v)$ the cosine of two shot features (public ViCLIP-L, eight frames per shot at a fixed stride, 224×224 without cropping). Let $\mathcal{P}^{\mathrm{GC}}_{jm} = \{(u, v) : u < v\}$ be all shot pairs after the near-duplicate filter, and $\mathcal{P}^{\mathrm{IC}}_{jm}$ and $\mathcal{P}^{\mathrm{CC}}_{jm}$ its pairs with $s_u = s_v$ and $s_u \neq s_v$. For $a \in \{\mathrm{GC}, \mathrm{IC}, \mathrm{CC}\}$,

$$M^a_{jm} = \frac{1}{|\mathcal{P}^a_{jm}|} \sum_{(u,v)\in\mathcal{P}^a_{jm}} \cos(x_u, x_v), \qquad |\mathcal{P}^a_{jm}| > 0. \tag{7}$$

Stories count equally; paired differences use the stories valid for both systems, and 95% intervals resample stories (10,000 draws). The *VBench code* variant uses VBench's ViCLIP checkpoint (InternVid-10M-FLT), segment-midpoint sampling, and preprocessing.

DINOv2, face identity, and face match. Both use eight uniformly spaced frames per shot at 832×480. DINOv2 (ViT-B/14, CLS token) averages normalized frame embeddings into a shot embedding and is scored like ViCLIP. Face identity uses RetinaFace and ArcFace (insightface `buffalo_l`): faces with score ≥ 0.6 and box side ≥ 32 px are grouped within a shot (cosine ≥ 0.4 to a centroid); groups seen in two or more frames are the shot's identities; a shot pair scores the largest cosine between their identities, and a film the mean over pairs where both shots have one. Shots with an identity: 87% (CANVAS), 63% (ViMax), 67% (MovieAgent). *Face match to seed portraits* (Table 6 and Section 4.4) uses the same detector and frames but compares each face with the closest of the story's seed portraits rather than with other shots of the film; a frame scores the mean of its faces' best cosines, a shot the mean over frames with a face, and a film the mean over such shots. It measures resemblance to fixed reference photographs, not consistency between shots, so its values (0.152 for the full ablation arm) are not comparable with face identity (0.398 for R1 in Table 2).

VBench quality. Aesthetic quality, imaging quality, subject consistency, and temporal flickering use VBench's code per shot, averaged per film.

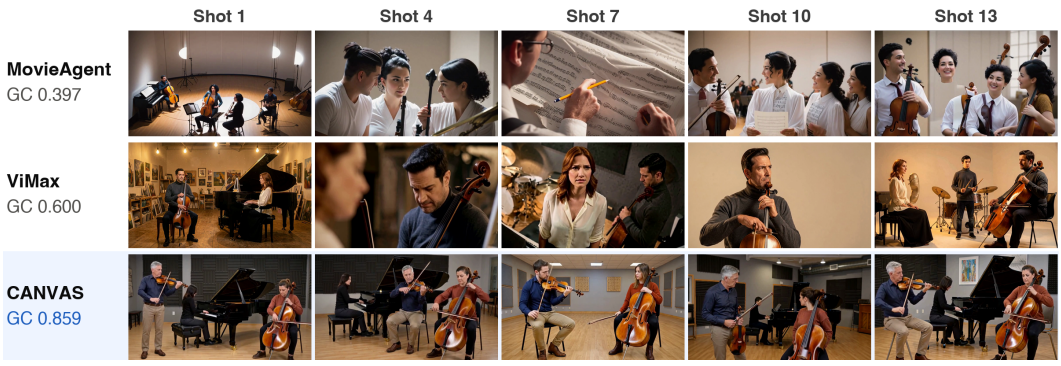


Figure 8: **The same story on the shared backend** (*The Final Run-Through*, the story at the median GC gap). Per shot, the frame whose face least matches the system's first shot. CANVAS keeps the three musicians and the room; ViMax's room and MovieAgent's players change.

J QUALITATIVE EXAMPLES

Figures are selected examples, not a prespecified sample.

K EXECUTION PROCEDURE AND AN EXAMPLE CONTRACT

Shot execution with bounded revision. Input: validated plan P, references $\mathcal{R}$, revision reserves r_i.

1. For each shot i in plan order, take h_i from the committed predecessor if $\sigma_i = 1$; otherwise render an identity-pack opening.
2. Compile the contract into call-local prompts.
3. For $k = 1, \ldots, 1 + r_i$: generate a take under the allocated steps; verify it; if it passes every gate, commit it and stop; otherwise revise the prompt toward the worst failed check.
4. If no take passed, commit the best-ranked one as a failed shot, keeping its failed verdicts.
5. Bind the committed take and its inputs to the execution record; only its last frame may condition a continuation.

Example contract. A bowl holding one vegetable batch sits on a counter beside a wok. The chef picks up the bowl with both hands, transfers the batch into the wok, returns the empty bowl to the counter, and releases it; a notebook stays on a dry support throughout. The batch moves from bowl to wok exactly once, and splitting pickup, transfer, and release makes hand occupancy explicit, so holding the bowl while writing with both hands is rejected before rendering.

L EXTENDED RELATED WORK

Long video generation. Video diffusion and transformer backbones synthesize short clips from text and visual conditions (Ho et al., 2022b; Blattmann et al., 2023a; Yang et al., 2024; Kong et al., 2024; Team Wan et al., 2025); streaming, hierarchical, and memory mechanisms extend their temporal scope (Yin et al., 2023b; Henschel et al., 2025; Lu et al., 2024; Zhou et al., 2026), and cross-image attention, storyboards, and multi-shot models add sequence structure (Zhou et al., 2024; Zhang et al., 2025b; Meng et al., 2025). All change the generator or its conditioning; CANVAS keeps the backend frozen and decides per shot which conditioning and settings it receives.

Agentic film production. VideoDirectorGPT plans scenes and entities before generation (Lin et al., 2023a), MovieAgent assigns hierarchical production roles (Wu et al., 2025), and others study collaborative and cinematic planning (He et al., 2024a; Xu et al., 2025; Huang et al., 2025; Wang et al., 2026). AniMaker searches over candidate clips (Shi et al., 2025). ViMax adds narrative retrieval, a camera-dependency graph for visual references, and best-of-k keyframe selection (Huang et al., 2026a). These systems are not context-blind; CANVAS differs in deciding, before rendering, which pixels a shot inherits (Equation 2), which object states must hold (Equation 3), and its denoising steps (Equation 4), instead of one global setting. GroundShot (Lai et al., 2026), closest to this work, is a training-free agentic framework that builds an entity-level visual memory from accepted shots, schedules the generation order of shots by their expected usefulness as entity references, and retrieves verified references before each shot; CANVAS instead fixes each shot's boundary, history, and steps in a plan compiled before rendering, in story order.

Feedback, inference compute, and evaluation. VideoAgent and NEWTON refine or replan generation with feedback and verifiers (Soni et al., 2024; Feng et al., 2026), following reflection-based agents (Yao et al., 2022; Shinn et al., 2023). Faster samplers and test-time scaling change the cost of one call (Song et al., 2020; Lu et al., 2022; Liu et al., 2025); CANVAS asks where to spend effort across a film's shots. VBench-style suites, ViCLIP, ArcFace, DINOv2, flow, and physics benchmarks supply the measurements we use (Huang et al., 2024; He et al., 2024b; Wang et al., 2023d; Deng et al., 2018; Oquab et al., 2023; Teed & Deng, 2020; Bansal et al., 2024).

Story and shot structure. Narrative planning before surface realization is long established in story generation, where an outline is produced first and expanded afterwards (Yao et al., 2018; Wang et al., 2024). Film-oriented systems inherit that separation and add production roles and camera language (Zhu et al., 2023; He et al., 2023; Zhao et al., 2024; Xiao et al., 2025; Huang et al., 2025). Storyboard anchoring and joint multi-shot modeling attack the same coherence problem from the representation side rather than the scheduling side (Zhang et al., 2025b; Meng et al., 2025; Zhou et al., 2026; Zhang et al., 2025a).

Backbones and conditioning. Our execution backend is interchangeable, and the surveyed families differ mainly in tokenizer, geometry, and training scale (Ho et al., 2022b; Singer et al., 2022; Ho et al., 2022a; Hong et al., 2022; Villegas et al., 2022; Blattmann et al., 2023b; Wang et al., 2023a; Chen et al., 2023a; 2024; Wang et al., 2023c; Zhang et al., 2023a; Girdhar et al., 2023; Gupta et al., 2023; Bar-Tal et al., 2024; Ma et al., 2024; Kondratyuk et al., 2023; Menapace et al., 2024; Blattmann et al., 2023a; Yang et al., 2024; Kong et al., 2024; Team Wan et al., 2025; HaCohen et al., 2024). Image-side ancestry and adapter-based control determine what a shot prompt can express (Rombach et al., 2021; Saharia et al., 2022; Podell et al., 2023; Zhang et al., 2023b; Ruiz et al., 2022; Guo et al., 2023; Wang et al., 2023b; Yin et al., 2023a; Xing et al., 2023a; Tu et al., 2024; Lin et al., 2025; 2023b). Long-context variants change how much history a call can absorb (Yin et al., 2023b; Ge et al., 2022; Harvey et al., 2022; Henschel et al., 2025; Chen et al., 2023b; Lu et al., 2024; Tewel et al., 2024), which is complementary to deciding how much history a shot should receive.

Cost, feedback, and tool use. Sampler and adapter work lowers the price of a single call (Song et al., 2020; Ho et al., 2020; Lu et al., 2022; Song et al., 2023; Xing et al., 2023b); test-time scaling and compute-optimal training ask instead how a fixed budget should be spent (Liu et al., 2025; Hoffmann et al., 2022). We take the second view at shot granularity. Agent scaffolds contribute the reflection and tool-routing patterns our loop reuses (Yao et al., 2022; Shinn et al., 2023; Schick et al., 2023; Huang et al., 2022; Du et al., 2023; Yuan et al., 2024; Tian et al., 2024; Li et al., 2025; OpenAI, 2023; Lin et al., 2024), and agentic decomposition has been applied to image synthesis and executable scenes as well (Li et al., 2026; Jiang et al., 2026; Yang et al., 2026; Cudlenco et al., 2026; Bai et al., 2025; He et al., 2024a; Zheng et al., 2025b; Xu et al., 2025; Wu et al., 2025; Wang et al., 2026). Physical plausibility supplies advisory diagnostics rather than gates (Wang et al., 2025; Kang et al., 2024; Motamed et al., 2025; Bansal et al., 2024; 2025).

Measurement. Benchmarks that decompose quality motivate our per-aspect reporting (Huang et al., 2024; Zheng et al., 2025a; Liu et al., 2024; 2023; He et al., 2024b; Rakheja et al., 2025; Meng et al., 2024), while feature-based scores set the limits we state (Wang et al., 2023d; Hessel et al., 2021; Zheng et al., 2023; Deng et al., 2018; Oquab et al., 2023; Teed & Deng, 2020).