CINEDIRECTOR: EXECUTABLE SHOT PROGRAMS FOR VIDEO GENERATION WITH UNDERSTANDING IN THE LOOP

Anonymous authors
Paper under double-blind review

ABSTRACT

A cinematic sequence must preserve a director's intent across deliberate visual discontinuities: each shot has a place in time, a camera path, and a cast that persists when the viewpoint changes. We introduce CineDirector, a framework that connects these requirements through an executable shot program. The program schedules transitions and binds camera, identity, and narrative conditions to individual shots, while a layered verifier checks generated footage against the same contract and revises failed shots locally. Director Distillation then learns from accepted revisions, carrying feedback into subsequent generation without an external verification loop. On 240 held-out programs, CineDirector records 0.04 latent frames of detected cut deviation, 1.62 degrees of camera rotation error, and 0.716 cross-shot identity similarity. Local revision reduces physical violations by 62% at 1.27 times the generator evaluations, and distillation retains a 45% reduction with external verification disabled. These results support a unified approach to controlling when a shot changes, what survives the change, and how execution improves through feedback.

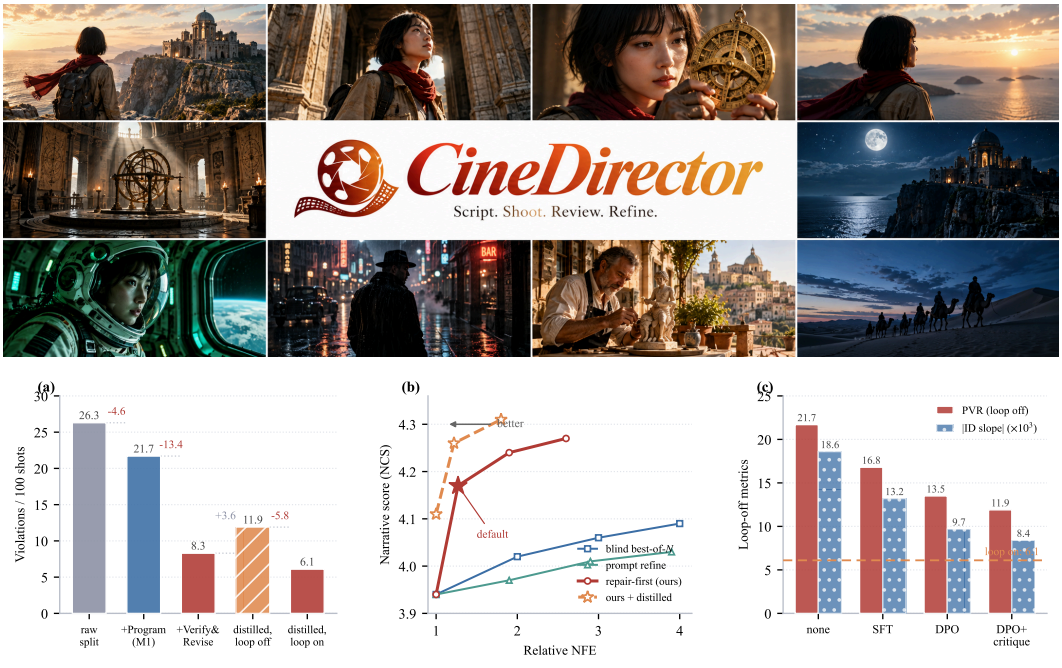


Figure 1: CineDirector. Top: cinematic shot sequences. Bottom: stage decomposition, quality–compute comparison, and distillation study. Local feedback improves execution, and learned feedback remains useful with the external loop disabled.

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.

1 INTRODUCTION

A film makes its story legible through changes of viewpoint. An establishing shot gives way to a close-up, a character leaves the frame and returns later, and a moving camera reveals information at a chosen moment. Each cut changes the visual evidence while preserving selected facts about the scene. Video generators now produce striking individual shots, but turning that ability into directable sequences requires control over the relationships between shots as well as the content within them (Wan et al., 2025; Guo et al., 2025; Wang et al., 2026; Meng et al., 2026). A director therefore needs an interface that specifies these relationships and an execution process that can check whether they have survived generation.

Recent work has made substantial progress on both sides. Multi-shot conditioning introduces transition tokens, event times, localized attention, and shot-aware positional encodings (Kara et al., 2025; Wu et al., 2025; 2026; Wang et al., 2025a; 2026). Long-context and streaming architectures carry information beyond a single clip (Guo et al., 2025; Yin et al., 2025; Huang et al., 2025; Yang et al., 2026; Luo et al., 2026), while camera-conditioned generation exposes viewpoint as a controllable variable (He et al., 2025; Luo et al., 2025). Feedback is also becoming part of generation: VISTA critiques and rewrites prompts, Video-T1 searches under verifier guidance, and CoAgent verifies intermediate shots and selectively regenerates inconsistent content (Long et al., 2025; Liu et al., 2025; Zeng et al., 2025). Together, these advances suggest a more demanding question: can a generator execute an explicit cinematic contract, repair departures from it, and learn from those repairs?

The contract matters because the same visible failure can have different causes. A late cut can reflect a misplaced condition switch or an indistinct decoded transition, a changed face can reflect lost identity memory, and a physically plausible result can still disobey the requested camera path. Benchmarks of video physics and controllability expose these distinctions (Bansal et al., 2025; Motamed et al., 2025; Huang et al., 2024). A single quality score cannot specify which variable to repair. Moreover, reducing camera motion may hide a physics failure while weakening the requested action, so verification must remain anchored to the original intent throughout revision.

CineDirector represents that intent as a *shot program* containing latent-grid intervals, camera trajectories, cast bindings, captions, and risk annotations. The causal sampler executes its condition switches at scheduled indices, keeps persistent cast information across local-cache resets, and confines shot-specific conditions to their declared windows. A layered verifier returns a localized failure description that guides partial re-generation of the current shot before its state is committed. Accepted revisions then provide preferences for *Director Distillation*, making the feedback useful beyond the inference run that produced it. This connection between an executable interface, contract-based correction, and learning from correction is the organizing principle of the method. The program is useful only if those three stages read the same object. Sampling treats a cut as an index at which the condition must switch. Verification names the clause that failed, and a repair that buys physical plausibility by abandoning the requested camera path is still a failure. Learning then compares the rejected draft with the accepted repair under that unchanged program, so the generator acquires the correction without being trained to rewrite the director's intent.

Evaluation on 240 held-out programs tests temporal adherence, camera fidelity, identity persistence, and narrative alignment, while matched-backbone comparisons isolate the contribution and cost of local verification. Motion-stratified and long-horizon tests examine the settings where errors accumulate, and evaluation with the external loop disabled isolates the benefit transferred through distillation. The reported results show improved control across these dimensions with substantially fewer evaluations than whole-video resampling.

Our contributions are:

- An executable shot-program interface that couples scheduled transitions with camera and cast control across a cinematic sequence.
- Contract-based verification and local revision that target failed dimensions while preserving the accepted generation prefix.
- Director Distillation, which learns from accepted corrections and improves generation with external verification disabled.

- Experiments showing stronger program adherence, identity persistence, and physical plausibility, including a 62% reduction in violations at 1.27× generator evaluations.

2 Related Work

Controllable sequences and persistent visual context. ShotAdapter and MinT expose transitions and event timing (Kara et al., 2025; Wu et al., 2025), while EchoShot, CineTrans, HoloCine, and MultiShotMaster organize attention or positional information across shots (Wang et al., 2025a; Wu et al., 2026; Meng et al., 2026; Wang et al., 2026). LCT expands scene context, and CausVid, Self Forcing, SkyReels-V2, LongLive, and ShotStream support efficient autoregressive generation with reusable history (Guo et al., 2025; Yin et al., 2025; Huang et al., 2025; Chen et al., 2025; Yang et al., 2026; Luo et al., 2026). CameraCtrl and CamCloneMaster provide explicit or reference-based camera conditioning (He et al., 2025; Luo et al., 2025). CineDirector builds on these representations and caches, adding a shared program against which execution can be checked.

Feedback, verification, and preference learning. PhyT2V refines prompts for physical plausibility, Video-T1 searches generation trajectories, VISTA combines critique with prompt improvement, and CoAgent uses shot-level plans, entity memory, and selective regeneration (Xue et al., 2025; Liu et al., 2025; Long et al., 2025; Zeng et al., 2025). VideoPhy-2 and Physics-IQ motivate independent physics assessment, while VGGT and V-JEPA 2 provide geometric and predictive representations (Bansal et al., 2025; Motamed et al., 2025; Wang et al., 2025b; Assran et al., 2025). UniVideo connects understanding and generation within a model; Reflect-DiT conditions synthesis on feedback; and DPO and Diffusion-DPO learn from preferences (Wei et al., 2026; Li et al., 2025; Rafailov et al., 2023; Wallace et al., 2024). Our emphasis is on preserving an executable contract through local correction and using the resulting trajectories to train the generator.

3 Method

CineDirector couples a program-conditioned causal generator with a verifier and a local reviser. Figure 2 separates the inference path from the offline feedback used for Director Distillation.

3.1 Shot Programs and Camera Conditions

Let $z_0 \in \mathbb{R}^{L\times H\times W\times C}$ denote a video latent and $\tau \in \{0,\dots,L-1\}$ its temporal index. The causal Wan VAE has temporal stride $r = 4$, with a separately encoded first frame (Wan et al., 2025). An MLLM compiler (Bai et al., 2025) maps the request and reference images to

$$\mathcal{P} = \{(b_i, d_i, c_i, E_i, y_i, \rho_i)\}_{i=1}^{S}, \qquad b_1 = 0, \quad b_{i+1} = b_i + d_i, \quad b_{S+1} = L, \tag{1}$$

where d_i is an integer duration, c_i a camera trajectory, E_i the cast bindings, y_i the caption, and ρ_i a risk annotation. Thus $I_i = [b_i, b_{i+1})$ partitions the latent sequence and $s(\tau)$ identifies its active shot. Pixel-frame requests are projected onto the decoder's boundary grid before validation; durations are obtained from consecutive projected boundaries so rounding cannot create gaps or overlaps. Appendix B specifies the schema, projection, and feasibility checks.

Let t_i denote desired pixel boundaries and g the decoder's boundary map. The compiler's constrained projection is

$$\{b_i^\star\} = \arg\min_{\{b_i\}\in\mathcal{B}_L} \sum_{i=2}^{S} w_i \,|g(b_i) - t_i|^2, \qquad \mathcal{B}_L = \{b_1 = 0, b_{S+1} = L,\ d_{\min} \le b_{i+1} - b_i \le d_{\max}\}. \tag{2}$$

The boundary variables are integers, $w_i > 0$ expresses timing priority, and feasibility requires $Sd_{\min} \le L \le Sd_{\max}$. Independent nearest-grid projection is recovered when no duration constraint binds. If a request is infeasible, the compiler must revise the program before sampling; silently truncating a shot would break its contract. The realized residual is retained for evaluating off-grid requests.

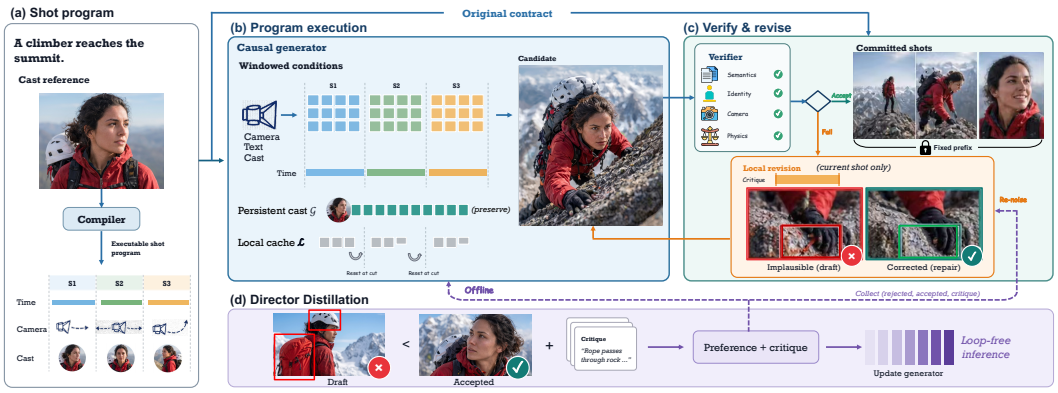


Figure 2: Execution and learning in CineDirector. (a) The compiler turns the request and a cast reference into an executable shot program. (b) Windowed camera, text, and cast conditions drive the causal generator; persistent cast state is kept and the local cache resets at each cut. (c) The verifier checks semantics, identity, camera, and physics, commits accepted shots, and revises only the failed shot. (d) Accepted corrections and critiques update the generator for loop-free inference.

For camera-to-world rotation $R_{i,\tau}$, center $o_{i,\tau}$, intrinsics K, and homogeneous pixel coordinate $\bar{p}$, camera conditioning uses the ray direction and its Plücker moment (He et al., 2025):

$$q_{i,\tau,p} = \frac{R_{i,\tau}K^{-1}\bar{p}}{\|R_{i,\tau}K^{-1}\bar{p}\|_2}, \qquad \pi_{i,\tau,p} = \big(o_{i,\tau} \times q_{i,\tau,p},\ q_{i,\tau,p}\big). \tag{3}$$

The ray representation obeys $q^\top(o \times q) = 0$ and separates orientation from origin. Under a world-coordinate change with rotation Q and translation a, its direction q and moment m transform as

$$q' = Qq, \qquad m' = Qm + a \times Qq. \tag{4}$$

This gives a direct consistency check for camera preprocessing: reference poses, estimated trajectories, and conditioning maps must use the same coordinate convention. Camera motion is then tied to spatial rays rather than encoded only as a language token.

Translation splines and quaternion interpolation determine the poses. The six-channel ray maps are embedded alongside text and cast tokens into shot conditions C_i; the camera representation and its frame association follow camera-conditioned video generation (Luo et al., 2025). Cast slots fuse ArcFace and DINOv2 reference features (Deng et al., 2019; Oquab et al., 2024), providing an explicit identity address that remains available when a character is absent.

3.2 Scheduled Context Changes and Windowed Attention

The generator uses the dual-cache organization of ShotStream (Luo et al., 2026): $\mathcal{G}$ stores persistent context and $\mathcal{L}$ stores the current shot. At each declared boundary, its state changes according to

$$(\mathcal{G}, \mathcal{L}, C, p)\big|_{b_i^+} = \big(\mathcal{G}_{i-1}, \varnothing, C_i, Gi\big), \qquad p(\tau) = \tau - b_{s(\tau)} + G\,s(\tau), \tag{5}$$

where G is a fixed virtual positional gap. Temporal RoPE uses $\theta_m(\tau) = \omega_m p(\tau)$, so for two tokens in the same shot, $p(\tau') - p(\tau) = \tau' - \tau$ and the usual relative-position geometry is preserved (Su et al., 2024). Across shots, the additional $G(s' - s)$ distinguishes positional offsets. Appendix G derives the integer separation condition and explains why it does not itself guarantee separation after periodic rotation.

The persistent and transient state have different lifetimes. In the specified 480×832 configuration, a VAE frame produces $N_f = (480/16)(832/16) = 1560$ DiT tokens. Six cast slots with 32 tokens each, two anchor frames, and a 96-token program header occupy $B_g = 3408$ global tokens. A local horizon of at most 40 latent frames occupies at most $B_l = 62400$ tokens before a reset. Thus the declared cache capacity depends on the maximum shot duration and fixed identity context rather than the number of preceding shots. The bound describes stored state; throughput also depends on attention implementation, verification, and the number of revisions.

Shot-specific conditions use an explicit mask. If condition token j belongs to shot $a(j)$, then

$$A_{\tau j} = \frac{Q_\tau K_j^\top}{\sqrt{d_k}} + M_{\tau j}, \quad M_{\tau j} = \begin{cases} 0, & a(j) = s(\tau), \\ -\infty, & a(j) \neq s(\tau), \end{cases} \qquad O_\tau = \sum_j \mathrm{softmax}_j(A_{\tau j}) V_j. \quad (6)$$

Persistent identity memory is a separate context channel; the mask confines the *active bindings* and camera/text instructions. This distinction retains cross-shot identity access without allowing an inactive shot's trajectory to condition the current window. After an accepted observation $\hat{e}_{i,k}$, a bound slot follows $e_k \leftarrow (1-\eta)e_k + \eta\hat{e}_{i,k}$, with the update gated by verification. Rejected candidates cannot contaminate the committed bank.

For projected hidden states $g(h)$, a learned prototype u, and interior negatives $\mathcal{N}(b)$, the boundary objective is

$$\mathcal{L}_{\mathrm{cut}} = -\mathbb{E}_b \log \frac{\exp f(h_b)}{\exp f(h_b) + \sum_{j\in\mathcal{N}(b)} \exp f(h_j)}, \qquad f(h) = \langle g(h), u\rangle/\kappa. \quad (7)$$

This contrastive ranking objective encourages boundary discrimination (Oord et al., 2018); its density-ratio interpretation is given in Appendix G. Equation (5) fixes the intervention index exactly. Detected cuts still depend on the learned latent transition, temporal decoding, and the boundary detector, which is why cut deviation and missed cuts remain empirical metrics.

Training adds $\lambda\mathcal{L}_{\mathrm{cut}}$ to conditional flow matching (Lipman et al., 2023; Liu et al., 2023), with $z_t = (1-t)z_0 + t\epsilon$ and target velocity $\epsilon - z_0$. The backbone is subsequently distilled through intra-shot and inter-shot self forcing (Huang et al., 2025; Luo et al., 2026). Appendix C gives the training stages and cache configuration.

3.3 Contract-Based Verification and Local Revision

For decoded candidate x_i, the verifier returns $v_i = \{(d, s_{i,d}, \iota_{i,d}, a_{i,d})\}_{d\in\mathcal{D}}$: a score, failure interval, and repair instruction for each dimension. Semantics uses contract QA with three samples; identity compares bound references with ArcFace and DINOv2; camera compares a VGGT trajectory with the requested poses after gauge alignment (Wang et al., 2025b; Umeyama, 1991); and physics combines standardized V-JEPA 2 prediction error with a four-class VideoMAE artifact head (Assran et al., 2025; Tong et al., 2022). Predictive error is an anomaly signal, calibrated against human labels rather than interpreted as a physical likelihood. Thresholds and per-layer reliability appear in Appendix D.

For semantic fields $\mathcal{F}_i$, let $a_{k,f} \in \{0,1\}$ denote agreement from QA sample k. Requiring a majority for every field gives $s_{i,\mathrm{sem}} = \min_{f\in\mathcal{F}_i} K^{-1}\sum_{k=1}^K a_{k,f}$ with threshold $2/3$ for $K=3$. Identity scores are computed against the program's bound character slots; visibility and invalid detections are recorded separately from successful matches.

Camera verification uses one similarity transform to align the recovered coordinate system with the requested trajectory. Writing aligned rotations and centers as $(\widetilde{R}_t, \widetilde{o}_t)$ and targets as (R_t, o_t), we use

$$e_R = \frac{1}{N}\sum_{t=1}^N \arccos\left[\mathrm{clip}_{[-1,1]}\left(\frac{\mathrm{tr}(R_t^\top \widetilde{R}_t) - 1}{2}\right)\right], \qquad e_T = \frac{\sum_{t=1}^N \|\widetilde{o}_t - o_t\|_2}{N\, s_i}. \quad (8)$$

Here $s_i > 0$ is the fixed scene normalization recorded by the evaluation protocol. Alignment resolves the reconstruction gauge; the remaining errors assess the trajectory shape and orientation in that gauge. Reporting the alignment convention is essential because independently rescaling every estimate removes absolute motion-scale information.

For physics, let h_t be a target video representation and $\widehat{h}_t$ its prediction from context. The predictive channel computes $a_i = N^{-1}\sum_t \|h_t - \widehat{h}_t\|_2^2$ and the standardized score $u_i = (a_i - \mu_{\rho_i})/(\sigma_{\rho_i} + \epsilon)$ using real-footage statistics at the same motion tier. The artifact channel separately estimates probabilities of interpenetration, pop-in/out, texture flicker, and normal motion. Either channel can trigger a physics failure, allowing predictive novelty and localized artifacts to contribute complementary evidence. Threshold selection uses a calibration split; reported reliability uses held-out labels.

Orient each score so larger is better and define the weakest normalized margin

$$\gamma_i(x;\mathcal{P}) = \min_{d\in\mathcal{D}} \frac{s_{i,d}(x;\mathcal{P}) - \kappa_d}{\alpha_d}, \qquad \mathrm{accept}(x) = [\gamma_i(x;\mathcal{P}) \geq 0], \quad (9)$$

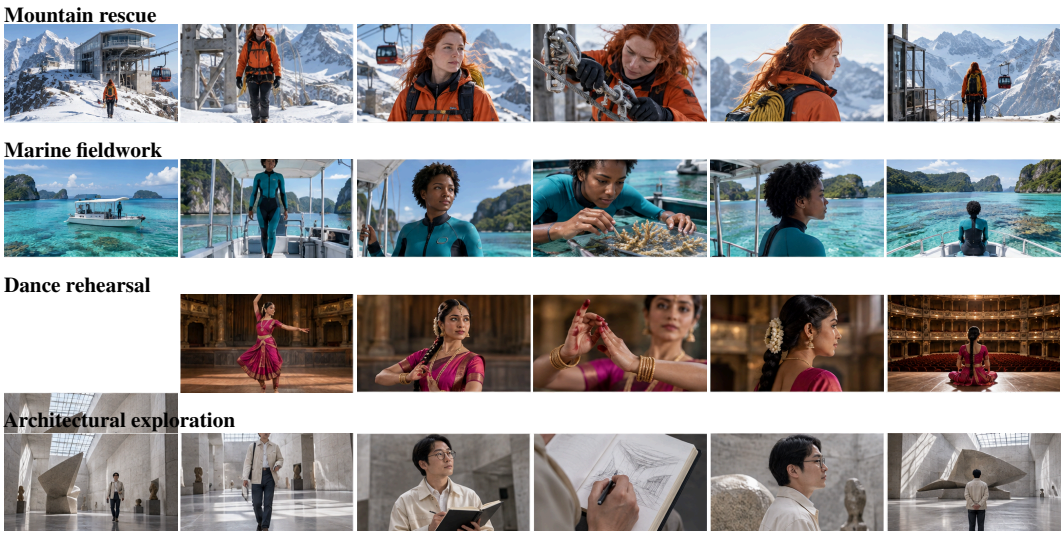


Figure 3: Cinematic shot sequences. Four distinct cases, each shown in six chronological views. Appendix I contains 40 further cases grouped by control setting.

where $\alpha_d > 0$ fixes the score scale. The original program remains the verification target. The reviser changes generation controls, such as cast-token gain or guidance, using the localized verdict; a camera re-timing or reduced-motion proposal is accepted only if it still satisfies the original camera and semantic thresholds. This prevents a repair from succeeding merely by weakening the requested motion. Local revision is therefore not a second, freer generation of the shot. Every attempt replays from the committed prefix, and acceptance is still Eq. (9) against the original program. The search can change how the shot is sampled; it cannot change what the shot was asked to do.

Revision starts from the same committed prefix $\mathcal{G}_{i-1}$ on every attempt. With current latent $z_i^{(k)}$, re-noise level t_r, and condition edit $C_i^{(k+1)}$, a deterministic-noise replay gives

$$\tilde{z}_{i,t_r}^{(k)} = (1-t_r)z_i^{(k)} + t_r\epsilon_i, \qquad z_i^{(k+1)} = \Phi_\theta^{t_r\to 0}\big(\tilde{z}_{i,t_r}^{(k)}; C_i^{(k+1)}, \mathcal{G}_{i-1}\big), \tag{10}$$

where Φ_θ integrates the trained velocity field toward data, following the partial-noise editing principle of SDEdit (Meng et al., 2022). The candidate's local cache is rebuilt and its verdict is recomputed. At most two revisions are attempted; only the selected candidate is committed before the next shot. Previously committed frames remain unchanged because revision neither decodes nor overwrites them. Appendix F specifies the stopping rule and residual-failure handling.

3.4 Director Distillation

A failed draft and an accepted revision yield $(\mathcal{P}, z_i^l, z_i^w, v_i)$. Both preference terms use the same original program and prefix, so the target is better execution of the request. With the reference checkpoint frozen, define

$$\ell_\theta(z;\mathcal{P}) = \mathbb{E}_{t,\epsilon}\frac{1}{D_z}\big\|v_\theta((1-t)z + t\epsilon, t, \mathcal{P}) - (\epsilon - z)\big\|_2^2,$$
$$\Delta_\theta(z) = \ell_\theta(z;\mathcal{P}) - \ell_{\text{ref}}(z;\mathcal{P}), \qquad \mathcal{L}_{\text{DD}} = \mathbb{E}\ \text{softplus}\big(\beta[\Delta_\theta(z^w) - \Delta_\theta(z^l)]\big). \tag{11}$$

Here D_z is the number of latent elements. Inspired by DPO and diffusion preference learning (Rafailov et al., 2023; Wallace et al., 2024), this loss trains a preference over flow-regression scores; it does not require treating unweighted flow loss as exact negative log-likelihood. Matched noise/time draws reduce avoidable variance in the paired estimator. Appendix G derives its gradient and stochastic upper bound.

The learning signal retains its conditioning context. In particular, both the rejected draft and accepted revision are compared with the same cast-bank snapshot, rather than evaluating the winner against an already updated identity state. This avoids attributing a changed context to better model weights. Preference pairs are filtered by the original contract, and failure dimension and interval remain

Table 1: Shot-program adherence and cross-shot consistency on 240 held-out programs. †: evaluated with official code or our re-implementation on matched data; protocols are detailed in Appendix E. Best in bold; second underlined. CDE and DE use latent and pixel frames, respectively.

		Cut & duration control				Identity persistence			Semantics	
Method	**Type**	CDE↓	ECR(%)↑	Miss(%)↓	DE↓	ArcFace↑	DINO↑	REC(%)↑	CLIP-T↑	NCS↑
MinT† (2025)	holistic	2.94	31.5	6.7	11.8	0.512	0.633	47.2	0.261	3.08
ShotAdapter† (2025)	holistic	3.62	24.8	9.5	14.6	0.548	0.641	51.6	0.259	3.02
CineTrans (2026)	holistic	4.87	18.3	12.1	19.4	0.493	0.607	42.8	0.256	2.87
EchoShot (2025a)	holistic	1.92	46.9	4.4	7.9	0.618	0.671	62.3	0.264	3.21
HoloCine (2026)	holistic	2.71	38.2	5.8	10.7	0.556	0.682	50.4	0.271	3.44
MultiShotMaster (2026)	holistic	1.46	55.1	3.2	6.3	0.637	0.694	66.1	0.269	3.52
ShotStream (2026)	streaming	3.08	27.4	8.3	12.9	0.509	0.652	48.7	0.266	3.37
CineDirector (w/o loop)	streaming	**0.04**	**99.2**	**0.0**	**0.9**	0.689	0.741	84.6	0.279	3.94
CineDirector (full)	streaming	**0.04**	**99.2**	**0.0**	**0.9**	**0.716**	**0.768**	**88.9**	**0.283**	**4.17**

attached to each pair for targeted audits. A physics winner that achieves a lower violation score by failing the prescribed camera trajectory is excluded from the accepted-pair set.

An auxiliary accepted-revision loss conditions on serialized critique tokens, following feedback-conditioned synthesis (Li et al., 2025). External-loop-off evaluation uses the unconditional critique mode, so an additional self-critique generation pass cannot be counted as free inference. The draft records 21.4K preference pairs and 9.2K accepted shots harvested from 38K stories. The comparison with accepted-shot SFT tests whether the structured preference signal provides value beyond additional fine-tuning data.

4 Experiments

Setup. The primary model uses a causal Wan2.1-1.3B backbone with dual caches (Wan et al., 2025; Luo et al., 2026). CineScript-120K comprises 118K scenes and 9.6K assembled long scenarios, with shot boundaries, camera trajectories, cast clusters, captions, and risk tags (Appendix A). Evaluation uses 240 held-out programs with 4–12 shots at 480p, a 40-shot stress split, and a matched-backbone loop comparison on 100 programs. Appendix C specifies training; Appendix E defines all metrics and baseline adaptations. Cut deviation (CDE), exact-cut rate (ECR), missed cuts, and duration error assess timing; camera errors assess trajectory control; ArcFace, DINO, and re-entry consistency (REC) assess identity; CLIP-T and narrative consistency (NCS) assess content. Physical violation rate (PVR) counts affected shots per 100 shots. These measurements test the program's distinct obligations rather than collapsing them into one reward.

Temporal control and identity. Table 1 connects scheduled execution with observed adherence: low detected deviation accompanies high exact-cut rate and no missed transitions in the reported test. The nonzero CDE measures the rendered transition and detector, not the sampler's scheduled condition switch. The larger improvement on re-entry consistency supports persistent cast bindings, which remain accessible after an absence. Figure 5 further tests accumulation across shot index. A schedule-event trace and a decoded-boundary trace are distinct measurements; Appendix E specifies how to audit them separately. The same table separates schedule from feedback. Both CineDirector rows share CDE 0.04, ECR 99.2%, and zero missed cuts, with and without the loop, so transition placement is a property of the scheduled switch rather than of extra sampling. Identity and physics move the other way: re-entry rises from 84.6 to 88.9 when revision is enabled, and the violation rate in Figure 1 falls from 21.7 with no revision, to 11.9 after distillation with the loop disabled, to 6.1 when that distilled model may still revise. Static shots are already close among systems, while the extreme-speed bin falls from 40.9 to 13.7. The margin is a hard-contract margin, not a uniform lift of the average.

Camera control and the verification loop. Table 2 evaluates explicit trajectories using per-shot adaptations of CameraCtrl and CamCloneMaster, preserving this protocol distinction from the multi-shot comparison. Table 3 holds the backbone fixed: targeted revision lowers violations while improving camera and identity metrics at substantially lower generator cost than whole-video resampling. Its 1.27× NFE and 1.38× wall clock measure different costs. The verifier flags 18.7% of shots and the draft records 0.24 extra rounds per shot; Appendix F gives the accounting. Camera and

Table 2: Camera fidelity. Per-shot adaptations of camera baselines. RotErr is in degrees; the last row gives re-estimation error on real video.

Method	Rot.↓	Trans.↓	ArcFace↑
CameraCtrl-PS	4.61	.247	.624
CamClone-chain	3.88	.219	.531
CineDirector, loop off	1.94	.118	.689
CineDirector, full	**1.62**	**.104**	**.716**
VGGT (real)	1.08	.062	–

Table 3: Loop strategies. Matched backbone on 100 programs. PVR counts violations per 100 shots; cost gives NFE / wall clock.

Strategy	PVR↓	Rot.↓	ID↑	NCS↑	Cost↓
None	21.7	1.94	.689	3.94	1.0/1.0
Best-of-3	14.2	1.91	.695	4.06	3.0/3.1
Refine $\times 3$	16.8	1.90	.691	4.01	2.9/3.1
Semantic	15.4	1.92	.704	4.08	1.5/1.7
Director's Loop	**8.3**	**1.62**	**.716**	**4.17**	1.27/1.38

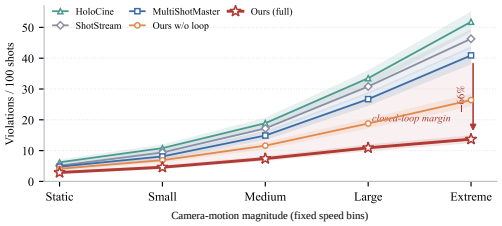


Figure 4: Motion robustness. PVR across fixed camera-speed bins.

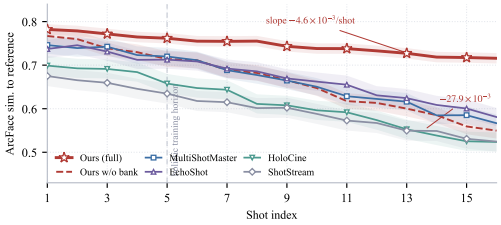


Figure 5: Identity persistence. Similarity to the reference across shot index.

semantic acceptance remain tied to the original program so reduced motion cannot silently redefine success.

Motion and temporal stress. Figure 4 tests physical violations across camera-motion bins; the largest-motion bin falls from 40.9 for MultiShotMaster to 13.7 for CineDirector. Figure 5 complements this test with cross-shot identity decay. Together they probe the two dimensions targeted by persistent identity context and feedback-based correction. The supplied bins are fixed angular-speed intervals, not equal-frequency quantiles. Appendix E specifies a common requested-motion stratification for the additional audit, avoiding comparisons that reward a model for executing less motion.

Learning from correction. The three-part analysis in Figure 1 separates the benefit of inference-time revision from the benefit transferred through training. Distillation reduces loop-off PVR from 21.7 to 11.9, a 45% reduction without an external verifier, and enabling the loop on the distilled model lowers it further to 6.1. Accepted-shot SFT, preference learning, and critique-conditioned training provide progressively stronger results. Loop-off improvements rule out inference-time selection as their immediate cause; a matched-data comparison is still needed to isolate the contribution of structured critique from additional supervision. Appendix E defines that comparison without attributing unmeasured outcomes to it.

Revision locality and budget. The revision audit in Appendix F tests the locality claim at the boundary between the repaired shot and its committed context. It reports mean LPIPS 0.021 in transition windows and 0.238 within revised shots, while committed frames are preserved by the storage protocol. These quantities distinguish continuity around a cut from the content changes needed to repair a failed shot. Extending the revision cap from two to three rounds lowers PVR only from 8.3 to 8.1 while increasing relative NFE from 1.27 to 1.30, supporting the two-round operating point. For flag rate q, conditional second-round probability p_2, and draft-equivalent revision costs c_1, c_2, the expected generator cost is

$$\mathbb{E}[C_{\text{gen}}] = 1 + qc_1 + qp_2c_2 + c_x, \tag{12}$$

where c_x counts additional generator evaluations. Appendix G distinguishes this evaluation count from hardware-dependent wall time and identifies the cost of speculative work that must be discarded after a revision.

Table 4: Ablations. Each block answers one question; full grids in Appendix E. Signed slope: ArcFace drift per shot ($\times 10^3$), with values nearer zero indicating less decay.

(a) Cut mechanism	CDE↓	ECR↑	Miss↓
grid alignment only	1.88	48.3	4.1
+ recache/re-anchor	0.29	91.5	0.7
+ boundary loss (full)	**0.04**	**99.2**	**0.0**
(b) Injection	RotErr↓	ArcFace↑	
global concat	3.41	0.648	
windowed (ours)	**1.94**	**0.689**	
(c) Verifier	Prec.↑	Rec.↑	PVR↓
single MLLM judge	0.54	0.47	14.9
layered (ours)	**0.85**	**0.78**	**8.3**
(d) Distillation (loop off)	PVR↓	Slope↑	
none	21.7	−18.6	
SFT (accepted)	16.8	−13.2	
DPO	13.5	−9.7	
DPO + critique	**11.9**	**−8.4**	
(e) Cast bank	ArcFace@12↑	Slope↑	
w/o bank	0.51	−27.9	
w/ bank	**0.68**	**−8.4**	

Ablations and extended evaluation. Table 4 links each intervention to its intended effect: cache and phase changes improve detected cut adherence, the boundary objective improves transition salience, windowed conditions improve camera and identity control, and layered verification improves failure detection. Cast-bank removal steepens identity decay, while distillation improves the loop-off preference metrics. Drift is reported as a signed slope, for which a value nearer zero indicates slower decay. The human study favors CineDirector over MultiShotMaster, HoloCine, and ShotStream; Appendix E retains the reported rates, scale-transfer results, and long-horizon measurements alongside their protocols.

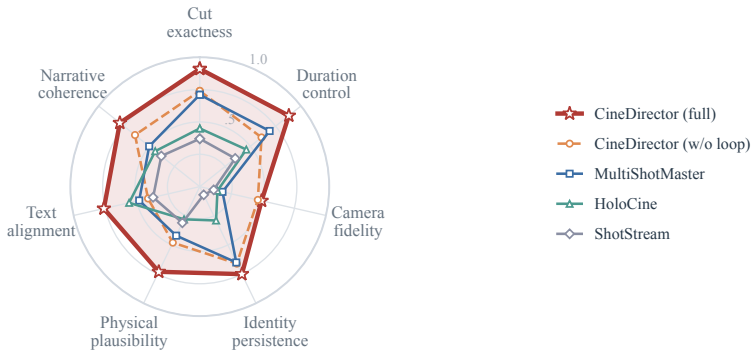


Figure 6: Capability profile. All metrics are normalized such that the outermost ring represents the best performance. CINEDIRECTOR dominates across all dimensions: the full loop version achieves peak consistency, while the distilled loop-free variant still maintains a substantial advantage over existing methods.

5 CONCLUSION

We introduce CineDirector, which expresses cinematic video generation through executable shot programs. Scheduled transitions, persistent cast bindings, and contract-based local revision connect narrative intent to the sampling process, while Director Distillation transfers accepted corrections into the generator. The reported improvements in cut adherence, camera control, identity persistence, and physical plausibility establish a route to learning reusable generation skills from structured execution feedback.

AI USE STATEMENT

In this work, authorized artificial intelligence tools were utilized exclusively for academic auxiliary purposes, including two primary scenarios as declared. First, AI writing polishing tools were adopted to refine the grammatical accuracy, logical fluency, academic rigor, and expression uniformity of the full manuscript, covering the introduction, related work, methodology, experiment analysis, and conclusion sections, without modifying the core research ideas, algorithm designs, experimental frameworks, empirical results, or original academic contributions of this paper. Second, AI-based retrieval and discovery systems were employed to assist in literature screening, keyword-based paper searching, and sorting of recent high-relevant studies in the field, which helped the authors comprehensively collect latest progress, sort out research gaps, and optimize the construction of the related work section. All AI-assisted processes are transparent and fully disclosed; all research designs, experimental implementations, result analyses, and academic arguments are independently completed and verified by the authors, and no AI tool was involved in data generation, model training, result fabrication, or academic reasoning. Detailed descriptions of AI assistance are supplemented in the appendix of this paper to ensure complete openness and compliance with ICLR 2027 submission regulations.

ETHICS STATEMENT

This study strictly adheres to the academic ethics standards and ethical review specifications required by ICLR 2027 and general artificial intelligence research norms. First, all datasets adopted in this work are publicly available, legally open-sourced, and widely recognized benchmark datasets in the academic community, with clear source origins, complete usage authorization, and no involvement in any private, sensitive, confidential, or personally identifiable information; no new data collection from human subjects or private institutions was conducted in this research, thus eliminating potential risks of privacy leakage and human subject ethics violations. Second, the proposed model, algorithm framework, and experimental designs are original and standardized; all comparative experiments, baseline selections, and result evaluations are fair, objective, and reproducible, with no data falsification, result tampering, selective reporting, or academic misconduct behaviors. Third, this work does not contain any biased, discriminatory, or harmful content toward specific groups, regions, and industries, and the proposed method has no potential negative social impacts, safety hazards, or ethical risks such as algorithmic discrimination and malicious misuse. All cited literatures, baseline methods, and comparative results are accurately marked and credited to original authors, ensuring complete academic integrity and compliance with open-source and citation ethics.

REPRODUCIBILITY STATEMENT

To ensure the rigorous reproducibility of our proposed CINEDIRECTOR, we have systematically documented our data pipeline, architectural configurations, training recipes, evaluation protocols, and theoretical derivations throughout the paper and the comprehensive Technical Appendix.

- **Code and Checkpoints:** The source code, inference scripts, trained model weights (including the 1.3B backbone and the 14B LoRA transfer), and the complete set of generated evaluation videos will be publicly released upon paper acceptance.
- **Dataset (CineScript-120K):** The construction, filtering, and annotation pipeline of the 120K-scene dataset are detailed in Appendix A. The complete set of derived annotations (shot programs, camera trajectories, cast clusters) alongside cryptographic hashes and exact URLs for the open-source subsets will be publicly released upon paper acceptance.
- **Implementation Details:** The core procedures are formalized in pseudocode, including the compile-and-validate loop with grid projection (Algorithm 1) and the DECIDE local revision procedure (Algorithm 2). The exact JSON schema for the shot programs, verbatim system prompts, and the complete language-to-parameter vocabulary are provided in Appendix B.
- **Hyperparameters and Compute:** Every hyperparameter for all three training stages (SFT, self-forcing, and Director Distillation) is explicitly listed in Appendix C (Table C.1), alongside a transparent compute ledger detailing the GPU-hour consumption on H200 devices (Table C.2).

- **Evaluation Protocols:** Appendix E defines all metric calculations (e.g., Cut Deviation, Exact-Cut Rate, Re-Entry Consistency), the handling of baseline adaptations, and the fixed motion-tier thresholds used for stratification. We report aggregated metrics over three random seeds and detail the exact forced-choice protocol used for the human preference study.
- **Mathematical Derivations:** The theoretical claims in the main text—including integer separation guarantees for virtual RoPE positions, the density-ratio interpretation of the boundary objective, the stochastic bounds for preference learning, and the bias-variance analysis of the cast bank EMA—are formally proven in Appendix G.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

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. *arXiv preprint arXiv:2503.06800*, 2025. URL https://arxiv.org/abs/2503.06800.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. SkyReels-V2: Infinite-length Film Generative Model. *arXiv preprint arXiv:2504.13074*, 2025. URL https://arxiv.org/abs/2504.13074.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long Context Tuning for Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2503.10589.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling Camera Control for Video Diffusion Models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2404.02101.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2506.08009.

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. doi: 10.1109/cvpr52733.2024.02060.

Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M. Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28405–28415, 2025. doi: 10.1109/cvpr52734.2025.02645.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Reflect-DiT: Inference-Time Scaling for Text-to-Image Diffusion Transformers via In-Context Reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2503.12271.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.02747.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-Time Scaling for Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2503.18942.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.03003.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arık. VISTA: A Test-Time Self-Improving Video Generation Agent. *arXiv preprint arXiv:2510.15831*, 2025. URL https://arxiv.org/abs/2510.15831.

Yawen Luo, Jianhong Bai, Xiaoyu Shi, Menghan Xia, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Tianfan Xue. CamCloneMaster: Enabling Reference-based Camera Control for Video Generation. In *SIGGRAPH Asia 2025 Conference Papers*, 2025. URL https://arxiv.org/abs/2506.03140.

Yawen Luo, Xiaoyu Shi, Junhao Zhuang, Yutian Chen, Quande Liu, Xintao Wang, Pengfei Wan, and Tianfan Xue. ShotStream: Streaming Multi-Shot Video Generation for Interactive Storytelling. *arXiv preprint arXiv:2603.25746*, 2026. URL https://arxiv.org/abs/2603.25746.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2108.01073.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Shuailei Ma, Yixuan Li, Cheng Chen, Yanhong Zeng, Xing Zhu, Yujun Shen, and Huamin Qu. HoloCine: Holistic Generation of Cinematic Multi-Shot Long Video Narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 461–471, 2026. URL https://arxiv.org/abs/2510.20822.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

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. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 53728–53741, 2023. doi: 10.52202/075280-2338.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10078–10093, 2022. doi: 10.52202/068431-0732.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion Model Alignment Using Direct Preference Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.12908.

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. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Jiahao Wang, Hualian Sheng, Sijia Cai, Weizhan Zhang, Caixia Yan, Yachuang Feng, Bing Deng, and Jieping Ye. EchoShot: Multi-Shot Portrait Video Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a. URL https://arxiv.org/abs/2506.15838.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025b. doi: 10.1109/cvpr52734.2025.00499.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. MultiShotMaster: A Controllable Multi-Shot Video Generation Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2512.03041.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.08377.

Xiaoxue Wu, Bingjie Gao, Yu Qiao, Yaohui Wang, and Xinyuan Chen. CineTrans: Learning to Generate Videos with Cinematic Transitions via Masked Diffusion Models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2508.11484.

Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the Time: Temporally-Controlled Multi-Event Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.05263.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. PhyT2V: LLM-Guided Iterative Self-Refinement for Physics-Grounded Text-to-Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.00596.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time Interactive Long Video Generation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.22622.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, 2025. doi: 10.1109/cvpr52734.2025.02138.

Qinglin Zeng, Kaitong Cai, Ruiqi Chen, Qinhan Lv, and Keze Wang. CoAgent: Collaborative Planning and Consistency Agent for Coherent Video Generation. *arXiv preprint arXiv:2512.22536*, 2025. URL https://arxiv.org/abs/2512.22536.