REFRAIN: LEARNING MULTI-SHOT CONSISTENCY FROM INCONSISTENT REFERENCES FOR A^2O VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A comic-book character, a photorealistic cityscape, an oil-painted coffee cup — such stylistically diverse references provide visual guidance across shots and frames, yet generating a coherent multi-shot video that unifies their visual styles while preserving subject identities and scene content remains challenging. Existing methods maintain cross-shot continuity through additional attention control or memory processing. However, they struggle to resolve disparate per-shot styles or handle reference–target duration mismatches within a single framework. To tackle this, we introduce **Refrain**, a unified model that learns cross-shot consistency through a self-supervised reconstruction task: reference clips from the same original video are deliberately perturbed with distinct visual edits, challenging the model to recover the unedited source footage under a shared style prompt. In this way, Refrain learns to suppress conflicting stylistic artifacts while faithfully retaining reference content and motion dynamics. Operating on the A^2O principle (**Any Reference, Any Duration, One Timeline**), Refrain maps image and video references onto a shared temporal canvas without auxiliary conditioning modules. For evaluation, experiments show that standard CLIP or DINO metrics frequently overlook stylistic clashes or misinterpret intended narrative cuts (such as changes in subjects or scenes) as inconsistency. We therefore propose **Shot-Debiased Consistency (SDC)**, which uses an extensive real-video bank to establish a baseline for style variation associated with content changes within shots. By discounting this baseline divergence, SDC evaluates cross-shot style shifts against real-world distributions, where smaller residual differences indicate higher stylistic continuity. We further construct **DurMix-300K**, a dataset pairing stylistically altered clips of varying lengths with their original video sequences, to train the model. Extensive experiments demonstrate that Refrain achieves state-of-the-art performance in cross-shot consistency, while excelling at shot timing, reference adherence, and text alignment.

1 INTRODUCTION

Generating coherent multi-shot videos from heterogeneous references (e.g., combining a photorealistic live-action clip with an animated reference) requires maintaining a unified visual style across shots while preserving underlying content and strict temporal boundaries (Figure 1). Existing frameworks use cross-reference attention, shot-specific masks, or visual memory to maintain continuity (Liu et al., 2025b; Meng et al., 2026; Zhang et al., 2025). Other methods support flexible condition placement or temporal control of reference subjects (Cai et al., 2026; Girish et al., 2026). We study the joint problem of resolving per-shot style conflicts and accommodating reference–target duration mismatches while preserving reference content and motion.

Therefore, we propose Refrain, a unified architecture that learns cross-shot style consistency through a self-supervised reconstruction task (Figure 2). The name Refrain evokes a recurring musical phrase, reflecting visual continuity across shot transitions, while its "Ref" prefix highlights reference-conditioned generation. Rather than relying on handcrafted consistency constraints, Refrain deliberately perturbs reference clips sampled from distinct shots of the same video with independent visual edits (e.g., style transfer, lighting variations, or weather shifts), introducing synthetic per-shot style

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.

conflicts. The model is then tasked with reconstructing the original target video under a shared style prompt. The generator receives the original shot descriptions and shared style prompt, without the edit instructions. Reconstructing the coherent target from independently edited references provides supervision for cross-shot style consistency. To support this supervision, we construct DurMix-300K, a large-scale dataset featuring over 300,000 reference conditions that pairs independently edited reference clips of arbitrary lengths with coherent target videos.

Refrain jointly synthesizes the entire multi-shot sequence in a single diffusion trajectory, using shot-aware temporal position offsets to maintain sharp cut transitions. To accommodate heterogeneous multimodal references on this unified timeline, Refrain operates on the A^2O principle: **Any Reference, Any Duration, One Timeline**. Specifically, reference images and variable-length video clips are mapped directly onto a shared temporal canvas, sparsely occupying their designated shot intervals via an occupancy mask without interpolation. By consolidating diverse inputs into a single sequence, a single backbone natively unifies text-to-video (T2V), video-to-video (V2V), text-image-to-video (TI2V), and composite (TIV2V) generation without auxiliary conditioning modules.

When evaluating cross-shot consistency of multi-shot videos, existing metrics like CLIP or DINO conflate legitimate narrative transitions with stylistic drift: they can overlook style conflicts and mistake content-feature changes caused by switching subjects or scenes for style inconsistency. We address this through a central question: *how much style variation is naturally expected when content changes across shots?* To answer this, we propose Shot-Debiased Consistency (SDC), a metric that calibrates the natural coupling between content shifts and style fluctuations using a real-video (~2 million video clips) bank. SDC estimates the style variation associated with cross-shot content changes, subtracts this baseline, and compares the residual against real-video samples with comparable content changes.

Experiments show that independent-edit training improves consistency under per-shot inconsistent test references while retaining reference adherence. Refrain achieves state-of-the-art overall cross-shot consistency, shot timing, text alignment, and reference adherence on the combined four-task benchmark. Our contributions:

- **Self-Supervised Reconstruction Learning.** We train on independently edited, stylistically inconsistent references paired with coherent targets. DurMix-300K provides these pairs with variable reference durations.
- **Unified Sparse Conditioning (A^2O).** One backbone handles T2V, V2V, TI2V, and TIV2V through sparse reference placement on a shared canvas, with no added modules.
- **Shot-Debiased Consistency.** SDC measures cross-shot style at comparable content change. Controlled edits and human judgments validate the measure, with a fixed real-video bank providing a common reference across generators.

2 Related Work

Multi-Shot and Reference-Conditioned Video Generation. Existing multi-shot and reference-guided generation frameworks typically maintain cross-shot continuity using auxiliary attention control, memory banks, or shot-specific masks (Deng et al., 2026; Girish et al., 2026; Meng et al., 2026; Zhang et al., 2025). Other methods support flexible visual conditioning and reference placement (Cai et al., 2026; Chen et al., 2026; Guo et al., 2024; Jiang et al., 2025). Prior data strategies also explore conditioning on unedited frames and identity-preserving augmentations (Chen et al., 2025; Yang et al., 2023; Zhou et al., 2026; Zhang et al., 2026). To address per-shot style conflicts and reference–target duration mismatches jointly, Refrain handles heterogeneous multimodal inputs on a single shared canvas under the A^2O principle and synthesizes the entire sequence in a single trajectory. By training on independently edited references paired with coherent targets, it learns to resolve cross-shot style conflicts without auxiliary conditioning networks.

Evaluating Cross-Shot Consistency. Standard video benchmarks evaluate visual quality using general perceptual metrics or local feature similarities across cuts (Huang et al., 2024; Zheng et al., 2024). While Content-Debiased FVD addresses semantic bias in overall video distributions (Ge et al., 2024) and CSD targets visual style representations (Somepalli et al., 2024), existing metrics conflate legitimate content transitions (such as changes in subjects, scenes, or viewpoints) with genuine



Figure 1: **Unified multi-shot generation with Refrain.** Each row shows a separate request with text, video, image, or joint conditioning. References vary in appearance and duration R, which can be shorter or longer than target duration D.

stylistic inconsistency. They fail to establish how much style variation is naturally expected across narrative cuts. In contrast, Shot-Debiased Consistency (SDC) decouples style drift from semantic progression by benchmarking cross-shot style separation against within-shot dynamics calibrated on an extensive real-video bank.

3 METHOD

Refrain learns cross-shot consistency through a self-supervised reconstruction task (Section 3.1). Independently edited references and their coherent source targets provide supervision for resolving per-shot style conflicts. The A^2O canvas (Section 3.2) represents image and video references of varying durations on the same target timeline, allowing one backbone to use this supervision across all four conditioning tasks (Figure 2).

To measure cross-shot visual style consistency without confounding content transitions or within-shot dynamics, we introduce Shot-Debiased Consistency (SDC) (Section 3.3).

3.1 LEARNING CROSS-SHOT CONSISTENCY THROUGH SELF-SUPERVISED RECONSTRUCTION

Reconstruction Target and Reference Edits. Let V be a coherent source video, y its original shot descriptions and shared style prompt, and $\mathcal{R} = \{R_i\}_{i=1}^{M}$ the selected references. An editing pipeline $\mathcal{E}$ transforms each reference using an instruction e_i from a distribution p_{edit} over the instruction library. Before quality filtering, shared and independent editing differ in how these instructions are

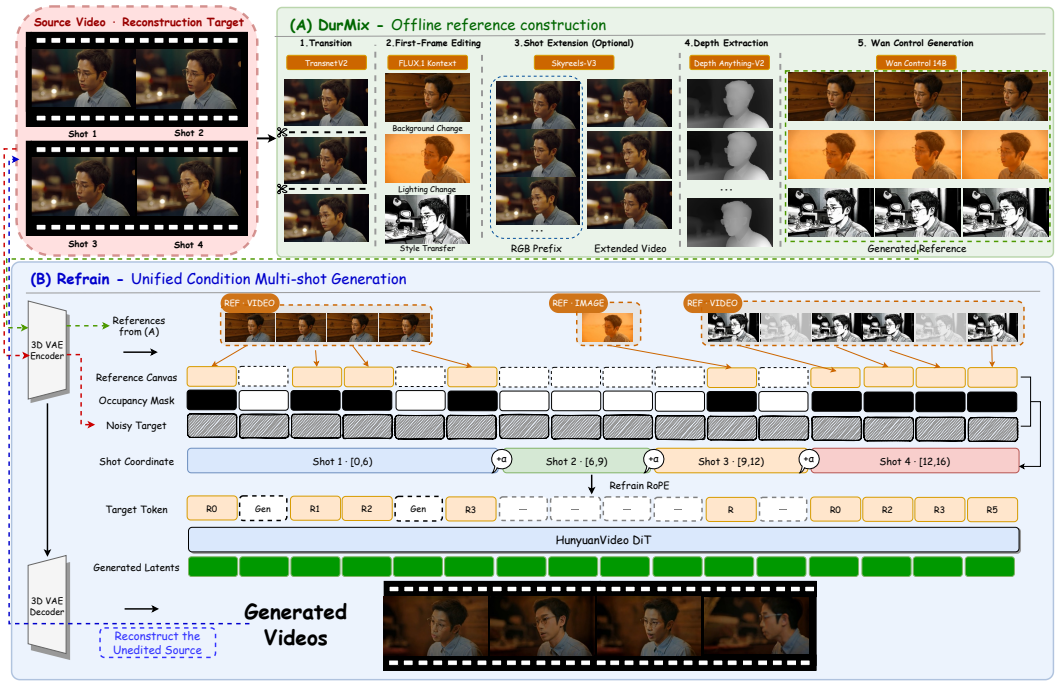


Figure 2: **Self-supervised reconstruction and unified conditioning.** (A) Reconstructing coherent source videos from independently edited references. (B) Joint multi-shot generation with sparse references on one timeline. Black indicates occupied mask positions ($m = 1$).

assigned:

$$\widetilde{R}_i = \mathcal{E}(R_i; e_i), \qquad \begin{cases} e_1 = \cdots = e_M = e, & e \sim p_{\text{edit}}, \quad \text{shared}, \\ e_1, \ldots, e_M \overset{\text{iid}}{\sim} p_{\text{edit}}, & \text{independent}. \end{cases} \tag{1}$$

Independent draws can assign, for example, a style change to one reference and a lighting change to another. Both strategies pair the edited references $\widetilde{\mathcal{R}}$ with the unchanged target V. Reconstructing that target encourages the model to use useful reference content while following the requested output style. Note that the generator receives y and $\widetilde{\mathcal{R}}$ without the edit instructions.

Training objective. We encode the original video as $z_0 = \text{Enc}(V)$ and form the standard flow-matching path from the clean latent to Gaussian noise (Lipman et al., 2023):

$$x_\tau = (1-\tau)z_0 + \tau\epsilon, \qquad u_\tau = \frac{\partial x_\tau}{\partial \tau} = \epsilon - z_0, \quad \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

Here $\tau \in [0,1]$ is sampled from the training noise-level distribution p_τ. Let $\mathcal{D}_{\text{ind}}$ contain the retained independent-edit training pairs, and let $\Phi(\widetilde{\mathcal{R}})$ be their canvas representation from Section 3.2. The model learns to predict the target velocity in Equation 2:

$$\mathcal{L}_{\text{ind}}(\theta) = \mathbb{E}_{\substack{(z_0, y, \widetilde{\mathcal{R}}) \sim \mathcal{D}_{\text{ind}} \\ \tau \sim p_\tau,\ \epsilon \sim \mathcal{N}(0,I)}} \left\| v_\theta\Big(x_\tau, \tau, y, \Phi(\widetilde{\mathcal{R}})\Big) - u_\tau \right\|_2^2. \tag{3}$$

Equation 3 supervises all target locations, including those conditioned on edited references. Shared-edit training replaces $\mathcal{D}_{\text{ind}}$ with $\mathcal{D}_{\text{shared}}$ and retains the same loss. Both strategies use the reconstruction objective without an auxiliary consistency loss. Section 4.3 tests the two strategies empirically.

3.2 Any Reference, Any Duration, One Timeline (A^2O)

We formulate a single unified interface supporting Text-to-Video (T2V), Video-to-Video (V2V), Text-Image-to-Video (TI2V), and Text-Image-Video-to-Video (TIV2V) without task-specific adapter modules or auxiliary networks. Given a target latent space $z_0 \in \mathbb{R}^{C \times T \times H \times W}$, a zero-initialized

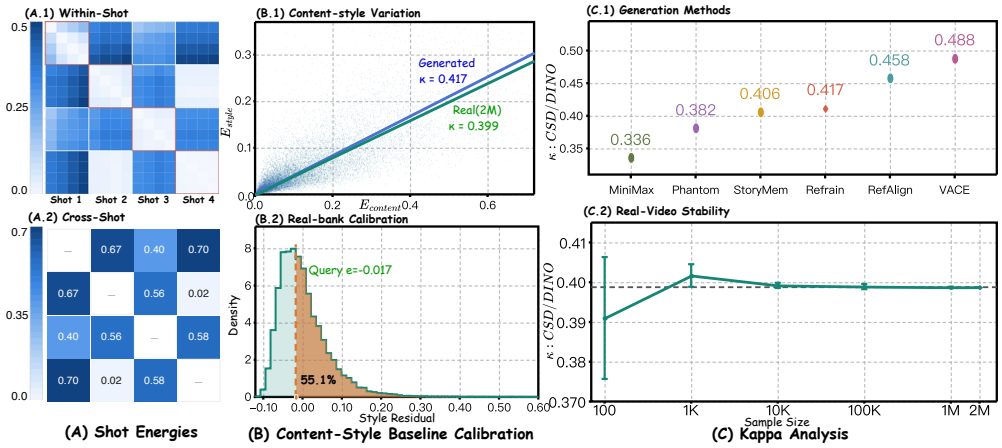


Figure 3: **SDC construction and motivation.** (A) Within- and cross-shot energies. (B) Content–style calibration. (C) Source-dependent κ and sample-size stability.

canvas latent $c \in \mathbb{R}^{C\times T\times H\times W}$ and a binary conditioning mask $m \in \{0,1\}^T$ track occupied temporal coordinates.

Sparse Mapping. For a reference sequence of length R assigned to a target interval $[d, d+D)$, we map reference latents $r_0, \ldots, r_{R-1}$ to destination indices via exact assignment:

$$\begin{cases} \pi(i) = d + \left\lfloor \dfrac{i(D-1)}{R-1} + \dfrac{1}{2} \right\rfloor, & c_{\pi(i)} = r_i, \quad m_{\pi(i)} = 1 \qquad (0 \le i < R \le D,\ R > 1), \\ \rho(j) = \left\lfloor \dfrac{j(R-1)}{D-1} + \dfrac{1}{2} \right\rfloor, & c_{d+j} = r_{\rho(j)}, \quad m_{d+j} = 1 \quad (0 \le j < D < R,\ D > 1). \end{cases} \tag{4}$$

For $1 < R \le D$, reference latents are assigned without temporal interpolation, leaving unoccupied positions at $m = 0$. For $R > D > 1$, source latents are subsampled with both endpoints retained. If $R = 1$ or $D = 1$, only r_0 is placed at c_d, with $m_d = 1$.

Shot-Aware Coordinates. To inform the model of shot transitions without breaking temporal self-attention across shots, we adopt the shot-aware temporal offsets of MultiShotMaster (Wang et al., 2026b) in Rotary Position Embeddings (RoPE) (Su et al., 2024). For a frame at temporal index t belonging to shot $s(t) \in \mathbb{N}_0$, its adjusted coordinate $p(t)$ is:

$$p(t) = t + \alpha s(t), \quad \alpha = 0.5. \tag{5}$$

3.3 Shot-Debiased Consistency

Cross-Shot Energy Distance. Evaluating raw frame features conflates cross-shot style drift with within-shot camera or object motion. We account for within-shot variation using an empirical energy contrast equivalent to squared maximum mean discrepancy (MMD) with a linear kernel (Székely & Rizzo, 2013; Gretton et al., 2012; Sejdinovic et al., 2013). Let $x^r_{i,a}$ be the normalized feature of frame a in shot i. We extract style features (CSD (Somepalli et al., 2024) and VGG AdaIN (Huang & Belongie, 2017), $r \in \{\text{CSD}, \text{VGG}\}$) and content features (DINOv2 CLS (Oquab et al., 2024), $r = c$). Defining frame distance $d^r(x, y) = 1 - \langle x, y\rangle$ and shot mean $\mu^r_i = \frac{1}{n_i}\sum_{a=1}^{n_i} x^r_{i,a}$, the empirical cross-shot energy contrast E^r_{ij} between shot i and shot j is (Figure 3(A)):

$$E^r_{ij} = 2 \cdot \operatorname*{mean}_{a,b} d^r(x^r_{i,a}, x^r_{j,b}) - \operatorname*{mean}_{a,a'} d^r(x^r_{i,a}, x^r_{i,a'}) - \operatorname*{mean}_{b,b'} d^r(x^r_{j,b}, x^r_{j,b'}) = \|\mu^r_i - \mu^r_j\|_2^2. \tag{6}$$

Content-Style Baseline Calibration. Frames within the same shot provide samples of natural variation without a shot transition. SDC uses these samples to establish a baseline for the style variation associated with content changes, then measures cross-shot style differences relative to this baseline. We construct a calibration bank from 2,144,201 real videos containing $6,462,692$ individual

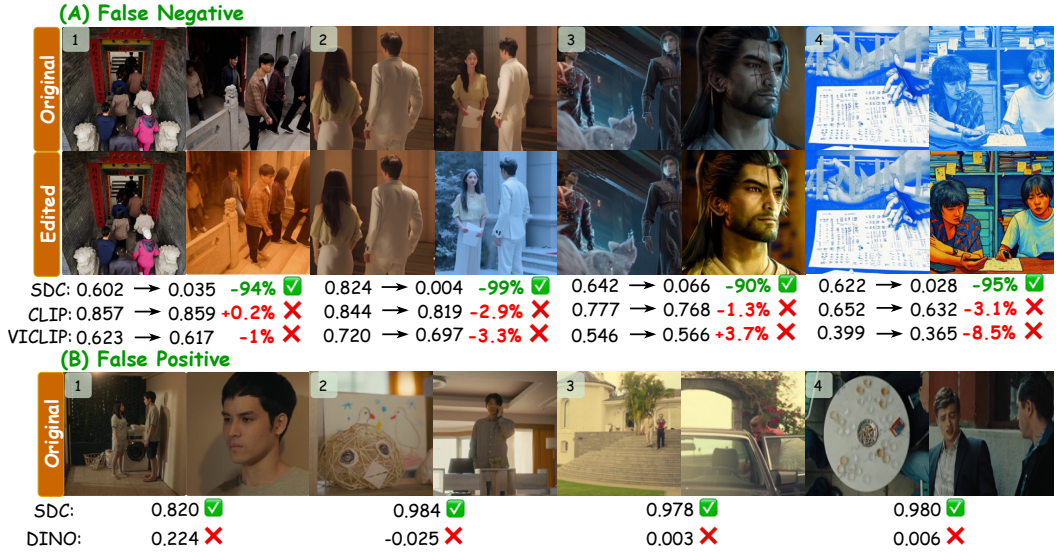


Figure 4: **Qualitative validation of SDC.** (A) SDC detects style edits missed by CLIP/ViCLIP. (B) SDC remains high across coherent scene cuts that lower DINO similarity.

shots. Each shot is split into balanced frame sub-groups A and B, yielding paired content and style split energies $C_b = \|\mu_A^c - \mu_B^c\|_2^2$ and $S_b^r = \|\mu_A^r - \mu_B^r\|_2^2$. We fit the expected content-to-style scaling slope $\hat{\kappa}_r$ via a through-origin Huber regression (Figure 3(B.1)):

$$\hat{\kappa}_r = \frac{\sum_b w_b a_b C_b S_b^r}{\sum_b w_b a_b C_b^2}, \quad a_b = \min\left(1, \frac{1.345\hat{\sigma}}{|S_b^r - \hat{\kappa}_r C_b|}\right), \tag{7}$$

where weights w_b balance video, shot, and split frequencies, and $\hat{\sigma}$ is updated via the weighted median absolute deviation of residuals. The fitted slope $\hat{\kappa}_r$ provides an empirical content–style calibration for the chosen real-video bank and feature representations, with stable estimates across sample sizes (Figure 3(C.2); Appendix D.1).

Conditional Residual Ranking. A video is scored by how much its style drifts beyond the expected baseline, compared against real videos with comparable content changes. For a query pair (i, j) with energies E_{ij}^c and E_{ij}^r, we subtract the expected content-induced shift to obtain the style residual:

$$e_{ij}^r = E_{ij}^r - \hat{\kappa}_r E_{ij}^c \tag{8}$$

Real-video residuals $z_b^r = S_b^r - \hat{\kappa}_r C_b$ are partitioned into 256 equal-mass content energy bins. The linear correction reduces content-related variation within each bin. For a query pair whose content energy E_{ij}^c falls into bin B, its style rank is computed via the weighted upper-tail survival percentile (Figure 3(B.2)):

$$\hat{G}_{B,r}(e) = \frac{\sum_{b \in B} w_b \left[\mathbb{I}(z_b^r > e) + \frac{1}{2}\mathbb{I}(z_b^r = e)\right]}{\sum_{b \in B} w_b}. \tag{9}$$

Linear interpolation between bin centers in content-percentile space gives the conditional survival estimate $\hat{G}_r(e \mid E_{ij}^c)$. The query rank is:

$$R_{ij}^r = \hat{G}_r(e_{ij}^r \mid E_{ij}^c) = \hat{G}_r\left(E_{ij}^r - \hat{\kappa}_r E_{ij}^c \;\middle|\; E_{ij}^c\right) \tag{10}$$

Finally, for a video V comprising $K \geq 2$ shots, we fuse the ranks from the CSD and VGG style heads in a fixed 2:1 ratio, and average over all $\binom{K}{2}$ shot pairs:

$$\mathrm{SDC}(V) = \frac{1}{\binom{K}{2}} \sum_{1 \leq i < j \leq K} \left(R_{ij}^{\mathrm{CSD}}\right)^{2/3} \left(R_{ij}^{\mathrm{VGG}}\right)^{1/3}. \tag{11}$$

At comparable content changes, smaller residual style differences receive higher real-bank ranks, yielding higher SDC scores and indicating greater cross-shot consistency.



Figure 5: **Multi-shot generation from per-shot inconsistent references.** Refrain follows the reference content and shot descriptions while generating a shared visual style across shots.

4 Experiments

4.1 Experimental Setup

Dataset and Task Allocation. We collect 2,150,758 web videos and annotate their shots and visual styles with Gemini-3 (Pichai et al., 2025) (example in Appendix A). We allocate 1,846,156 videos (85.84%) to T2V training, 297,192 (13.82%) to reference editing, and 7,410 (0.34%) to testing. We construct 300 evaluation requests from the test split across T2V, V2V, TI2V, and TIV2V. We construct DurMix-300K containing over 300,000 paired conditions with variable durations. For each selected reference shot, its initial frame is edited via FLUX.1 Kontext (Black Forest Labs et al., 2025) under one of 61 fine-grained edit instructions. DurMix combines style edits with attribute perturbations. Style edits create per-shot appearance conflicts for learning cross-shot consistency, while changes to subject appearance, spatial relations, and motion provide additional perturbations for reconstruction. To decouple reference duration from target length, clips are either cropped or extended with SkyReels-V3 (Li et al., 2026). A depth-conditioned Wan-2.1 model (Team Wan et al., 2025), guided by Depth Anything v2 (Yang et al., 2024), propagates the edited frame through the reference clip using depth-conditioned video generation. Finally, Qwen3-VL-235B (Bai et al., 2025) and TransNetV2 (Souček & Lokoč, 2024) filter out clips with structural degradation or unintended cuts (Appendix A).

Implementation Details. Refrain uses an 8.49B parameter HunyuanVideo-1.5 DiT architecture (Wu et al., 2025) as its underlying backbone. The model is first pre-trained for 19,000 steps on pure T2V tasks using 256 NVIDIA H800 GPUs (around 5 days), followed by 26,000 reference fine-tuning

Table 1: **Unified multi-shot generation.** Bold/underlined: best/second-best; ✗: unreported or inapplicable. Parameters include all experts and unmerged adapters. FLOPs are analytical denoising estimates for a 480×832, 241-frame, three-shot video with three 49-frame references under each method's sampling schedule and reference adapter, excluding encoders and VAE decoding. SDC uses a fixed real-video bank. Metric definitions: Appendix C.

Method	#Params (B)	FLOPs (P) ↓	Generation and Alignment				Inter-Shot ↑			Reference ↑		
			MAE s ↓	CLIP-T ↑	TA ↑	SDC ↑	Semantic	Subject	Scene	Semantic	Subject	Background
HunyuanVideo1.5	8.49	**324.4**	0.802	0.212	-0.386	0.090	0.566	0.439	0.478	✗	✗	✗
Phantom	14.29	648.7	✗	0.203	-0.675	0.130	0.491	0.345	0.385	0.556	0.450	0.376
RefAlign	14.29	807.7	✗	0.204	-0.530	0.125	0.493	0.341	0.361	0.614	0.469	0.448
StoryMem	29.80	<u>493.2</u>	✗	0.197	-0.679	0.175	<u>0.708</u>	0.489	**0.531**	0.478	0.412	0.388
VACE	17.34	608.9	✗	<u>0.217</u>	-0.336	0.108	0.492	0.362	0.361	✗	✗	✗
MiniMax-H3	33.12	550.7	<u>0.242</u>	0.216	<u>-0.267</u>	<u>0.295</u>	0.674	**0.516**	0.476	<u>0.686</u>	<u>0.559</u>	<u>0.541</u>
Refrain (ours)	8.49	**324.4**	**0.199**	**0.222**	**-0.221**	**0.315**	**0.710**	<u>0.494</u>	<u>0.491</u>	**0.737**	**0.671**	**0.674**

steps using 16 H800 GPUs (around 2 days): 17,000 steps with references matched to target durations, then 9,000 steps with variable-duration references. Text-only tasks and examples without selected references maintain T2V training. Condition dropout (rate 0.2) retains at least one active reference.

Baselines and Evaluation Metrics. We compare Refrain with representative multi-shot and reference-conditioned generation methods, including StoryMem (Zhang et al., 2025), Phantom (Liu et al., 2025b), RefAlign (Wang et al., 2026a), VACE (Jiang et al., 2025), and MiniMax-H3 (MiniMax, 2026). Cross-shot style consistency is evaluated using our proposed SDC. Transition timing accuracy is measured via Transition Mean Absolute Error (MAE). Text alignment is scored using CLIP-T and VideoAlign TA (Liu et al., 2025a). Inter-shot consistency is assessed across semantics, subjects, and scenes. Reference fidelity is measured by semantic, subject, and background consistency.

4.2 Results

As shown in Table 1, Refrain consistently outperforms baselines in visual consistency, reference adherence, and temporal precision.

Style Consistency and Reference Fidelity. Under test references with deliberate style conflicts, Refrain achieves the highest style consistency (SDC 0.315 vs. MiniMax-H3's 0.295). Reference Semantic reaches 0.737, ahead of MiniMax-H3's 0.686, alongside leading reference subject (0.671) and background (0.674) scores. These results support cross-shot style consistency with strong reference correspondence (Figure 5).

Transition Accuracy and Text Alignment. Refrain jointly generates all shots in a single diffusion trajectory and achieves a transition MAE of 0.199s. Meanwhile, leading VideoAlign TA (-0.221) and CLIP-T (0.222) scores demonstrate strong adherence to text prompts alongside visual harmonization.

4.3 Ablation Studies

Impact of Inconsistent Reference Supervision. We compare independent and shared reference-edit instructions under the same training budget (Equation 1). On independently edited test references, independent-edit training raises SDC from 0.301 to 0.315. With shared-edit test references, independent-edit training reduces SDC from 0.304 to 0.299 (Table 3(a)). The gain therefore occurs when the test references contain per-shot inconsistency.

Table 2: **Pure T2V generation.**

Training	SDC ↑
T2V pretraining	0.253
Independent edits	**0.303**

We also evaluate the same text-only requests before and after independent-edit fine-tuning, without visual references at inference. SDC increases from 0.253 after T2V pretraining to 0.303 after fine-tuning (Table 2), showing a consistency gain beyond reference-conditioned generation.

Shot Coordinates and Reference Placement. Table 3(b) verifies the core components of the A^2O canvas. Incorporating shot-aware RoPE offsets is vital for boundary precision: removing these offsets increases transition error from 0.199s to 0.281s. Furthermore, replacing our sparse latent placement

Table 3: **Ablation studies.** (a) Reference editing. (b) Shot coordinates and placement.

(a) Reference-editing supervision

Training edits	Test edits	SDC ↑
Independent	Independent	**0.315**
Independent	Shared	0.299
Shared	Independent	0.301
Shared	Shared	0.304

(b) Shot coordinates and placement

Coordinates	Placement	MAE s ↓	SDC ↑
Shot-aware	**Sparse**	**0.199**	**0.315**
Shot-aware	Resample	0.233	0.310
None	Sparse	0.281	0.304
None	Resample	0.293	0.289

Table 4: **Style and content controls.** (a) Single-shot style edits (108,909 videos). (b) Shared/independent edits (72,570 videos). Means and detection at 5% control false positives. (c) Real videos.

(a) Single-shot style edits

Metric	Score	Detect. ↑
CLIP	$0.836 \rightarrow 0.711$	48.6%
ViCLIP	$0.609 \rightarrow 0.426$	38.6%
DINO	$0.385 \rightarrow 0.297$	10.0%
SDC	$0.346 \rightarrow 0.047$	**78.0%**

(b) Independent reference edits

Metric	Score	Detect. ↑
CLIP	$0.771 \rightarrow 0.704$	15.2%
ViCLIP	$0.560 \rightarrow 0.420$	18.9%
DINO	$0.408 \rightarrow 0.312$	10.8%
SDC	$0.302 \rightarrow 0.078$	**50.9%**

(c) Real-video content changes

Subset	DINO	SDC
All (2.14M)	0.383	0.337
DINO < 0.25	0.170	0.339
Lowest DINO 10%	0.117	0.322
DINO < 0.10	0.059	0.287

with temporal frame resampling worsens transition error to 0.233s while reducing SDC from 0.315 to 0.310. Sparse placement avoids latent interpolation. Appendix H.1 shows how Refrain adapts reference actions to the requested shot duration while preserving natural motion pacing and a shared visual style.

4.4 Validating SDC

Style and Content Changes. Table 4 extends Figure 4 to large-scale controls. At 5% false positives, SDC detects 78.0% of single-shot style edits, versus 48.6% for CLIP and 38.6% for ViCLIP. It also better distinguishes independent from shared reference edits when both groups are edited. Among 2.14M real videos, the low-DINO subset (< 0.25) retains mean SDC 0.339 versus 0.337 overall, showing stable consistency scores despite substantial content changes. Appendix D compares SDC with CSD/VGG similarity and tests $\kappa = 0$ (Tables 10–11).

Calibration Stability and Source Variation. The real-video content–style slope remains close to 0.399 across calibration sizes from 100 to 2,144,201 videos (Figure 3(C.2)). Generated-source slopes range from 0.336 to 0.488 and overlap the real-video estimate (Figure 3(C.1)). These results support a stable real-video calibration (Appendix D.1).

Human Agreement. Six professional raters assessed cross-shot style consistency in 120 clips: 80 Refrain-generated, 20 real, and 20 spliced from edited segments. A four-of-six consensus identifies 104 consistent and eight inconsistent clips, with eight unresolved. SDC separates the two human-labeled classes with an AUC of 0.969 (Table 5), showing agreement with human judgments of style consistency. Appendix E gives the protocol and score distributions.

Table 5: **Human agreement.**

Statistic	SDC
Consistency AUC ↑	0.969
95% interval	[0.931, 0.995]
Permutation p	1.0×10^{-5}

5 Conclusion

Refrain learns cross-shot consistency through the self-supervised reconstruction task on DurMix-300K. Its shared timeline unifies text, image, and video conditioning without added modules. SDC evaluates the resulting consistency against natural content–style variation in real videos. Experiments show improved consistency and reference adherence under per-shot inconsistent inputs, with consistency gains extending to pure T2V generation.

AI USE STATEMENT

Large language models assisted with writing and code development, including analysis scripts. The authors are responsible for the manuscript and reported results. Gemini-3 provides video annotations, and Qwen3-VL provides data-quality filtering (Section 3.1 and Appendix A).

REPRODUCIBILITY STATEMENT

The supplement specifies data construction, training and inference settings, evaluation sets, metrics, and statistical procedures.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Minghong Cai, Qiulin Wang, Zongli Ye, Wenze Liu, Quande Liu, Weicai Ye, Xintao Wang, Pengfei Wan, Kun Gai, and Xiangyu Yue. VideoCanvas: Unified video completion from arbitrary spatiotemporal patches via in-context conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4475–4485, 2026.

Junyi Chen, Tong He, Zhoujie Fu, Pengfei Wan, Kun Gai, and Weicai Ye. VINO: A unified visual generator with interleaved OmniModal context. *arXiv preprint arXiv:2601.02358*, 2026.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6099–6110, 2025.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked guidance for any-reference video generation with subject disentanglement. In *International Conference on Learning Representations (ICLR)*, 2026.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in Fréchet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7277–7288, 2024.

Sharath Girish, Viacheslav Ivanov, Tsai-Shien Chen, Hao Chen, Aliaksandr Siarohin, and Sergey Tulyakov. AlcheMinT: Fine-grained temporal control for multi-reference consistent video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23227–23237, 2026.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL `https://www.jmlr.org/papers/v13/gretton12a.html`.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. SparseCtrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision (ECCV)*, pp. 330–348, 2024.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1501–1510, 2017.

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.

Peter J. Huber. Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821, 1973. doi: 10.1214/aos/1176342503.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Debang Li, Zhengcong Fei, Tuanhui Li, Yikun Dou, Zheng Chen, Jiangping Yang, Mingyuan Fan, Jingtao Xu, Jiahua Wang, Baoxuan Gu, Mingshan Chang, Wenjing Cai, Yuqiang Xie, Binjie Mao, Youqiang Zhang, Nuo Pang, Hao Zhang, Yuzhe Jin, Zhiheng Xu, Dixuan Lin, Guibin Chen, and Yahui Zhou. SkyReels-V3 technique report. *arXiv preprint arXiv:2601.17323*, 2026.

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.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82155–82192. Curran Associates, Inc., 2025a.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961, 2025b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024.

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.

MiniMax. MiniMax-H3. Hugging Face, official model card, 2026. URL `https://huggingface.co/MiniMaxAI/MiniMax-H3`. Accessed September 25, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. 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://openreview.net/forum?id=a68SUt6zFt.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. A new era of intelligence with Gemini 3. Google, official model announcement, 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3/. Accessed September 25, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi: 10.1214/13-AOS1140.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Investigating style similarity in diffusion models. In *European Conference on Computer Vision (ECCV)*, pp. 143–160, 2024.

Tomáš Souček and Jakub Lokoč. TransNet V2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11218–11221. ACM, 2024.

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

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

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.

Lei Wang, Yuxin Song, Ge Wu, Haocheng Feng, Hang Zhou, Jingdong Wang, Yaxing Wang, and Jian Yang. RefAlign: Representation alignment for reference-to-video generation. In *European Conference on Computer Vision (ECCV)*, pp. 647–664, 2026a.

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)*, pp. 16268–16278, 2026b.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, Linus, Patrol, Peizhen Zhang, Peng Chen, Penghao Zhao, Qi Tian, Songtao Liu, Weijie Kong, Weiyan Wang, Xiao He, Xin Li, Xinchi Deng, Xuefei Zhe, Yang Li, Yanxin Long, Yuanbo Peng, Yue Wu, Yuhong Liu, Zhenyu Wang, Zuozhuo Dai, Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou, and Zhao Zhong. HunyuanVideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18381–18391, 2023.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911. Curran Associates, Inc., 2024.

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

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. *arXiv preprint arXiv:2512.19539*, 2025.

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. *arXiv preprint arXiv:2412.02259*, 2024.

Zixiang Zhou, Zhentao Yu, Yifeng Ma, Hongmei Wang, Wenqing Yu, Cong Wang, Zilin Yang, Rui Chen, Jiarong Ou, Yezhou Liu, Yuan Zhou, and Qinglin Lu. Aura: Consistent multi-subject video generation via VLM-grounded semantic alignment. *arXiv preprint arXiv:2607.04311*, 2026.

A Data Construction

Source videos and annotations. We collect 2,150,758 web videos and use Gemini-3 to annotate shot boundaries, shot descriptions, entities, and visual style. The split assigns 1,846,156 videos (85.84%) to T2V training, 297,192 (13.82%) to reference-editing construction, and 7,410 (0.34%) to testing. The 300 evaluation requests are drawn from the test allocation. Original-video annotations provide the shot descriptions and shared style prompt. Table 6 presents an example annotation.

Table 6: Example Gemini-3 annotation.

Field	Annotation
Style	Realistic, cinematic thriller style.
Environment	A dimly lit urban street at night, lined with parked cars and trees. The setting suggests a residential or commercial area after hours.
Shot 1, 0.0–2.7 s	A woman (<ID:1>) holding a handgun (<ID:2>) looks over her shoulder and speaks.
Shot 2, 2.7–4.2 s	A man walks away from the camera towards a motorcycle on a dark street.

DurMix-300K pairs per-shot inconsistent references with coherent original videos to supervise reconstruction with a shared output style. Its construction plan contains 508,041 reference conditions, with 495,634 assigned to training and 12,407 to testing. The validated inventory contains 309,352

conditions: 200,265 shorter and 109,087 longer than their target intervals. These conditions belong to 223,850 source videos. The dataset's 300K designation refers to the reference-condition count. Among 76,532 videos with multiple video references, 80.03% use different edit categories and 98.49% use different instructions.

Editing and quality filtering. To vary reference appearance and scene attributes, we sample from 61 instructions in eight categories (Table 7). Each selected frame is edited and propagated through its reference clip as described in Section 3.1. Qwen3-VL-235B evaluates preservation of the reference, edit quality, and temporal validity. TransNetV2 rejects transitions inside intended single-shot controls. For motion and spatial-relation edits, filtering evaluates reconstruction suitability while accommodating the intended attribute changes.

Table 7: Editing categories in DurMix-300K. Counts refer to validated reference conditions.

Category	Instructions	Conditions
Style transfer	20	125,457
Subject appearance	8	31,655
Object material or colour	7	24,544
Lighting and weather	6	31,175
Background retexturing	6	28,718
Spatial relation	6	21,266
Camera or lens filtering	4	22,672
Motion change	4	23,865
Total	61	309,352

Roles of reference edits. All edits modify the reference inputs while leaving the reconstruction target unchanged. Style transfer, camera filtering, and lighting edits primarily introduce appearance differences across shots. Subject-appearance, spatial-relation, and motion edits additionally perturb attributes of the source content. Attribute perturbations train reconstruction from partially altered references. Recovery of an edited attribute depends on the information available in the remaining references and text. Reference preservation therefore concerns the source information retained after editing, rather than every attribute introduced by an edit. For example, a clothing or action change in a training reference does not change the clothing or action required by the target. During training, the original video and its annotations define the desired output, and edited references provide additional visual information.

Duration construction. Reference duration varies while the selected reference frame, control start, and generation target remain fixed. To avoid including a nearby transition, we remove exactly one latent step, or four pixel frames, from the retained real-control endpoint. Existing short conditions are cropped directly in latent space. After removing any legacy extension, their endpoint is $\min(F_{\text{planned}}, F_{\text{real}}) - 4$ pixel frames. Here F_{planned} is the assigned reference length and F_{real} is the available real-control length, both in pixel frames. For RGB extension, SkyReels-V3 fills the gap between the tail-trimmed real control and the assigned reference duration. The completed sequence preserves the exact real prefix. We re-estimate depth over that sequence before Wan Control propagation and latent encoding, retaining the completed sequence length. Extension supplies footage needed to construct a reference while preserving the requested output duration. The resulting reference can remain shorter or longer than its generation target.

Construction resources. Generation of DurMix-300K consumed an estimated 126 hours on 864 H20 GPUs. TransNetV2 filtering used 128 H20 GPUs for approximately 44 minutes. Qwen3-VL-235B, deployed on 128 H20 GPUs, performed screening and rechecking in approximately 16.4 hours.

B Training and Inference Settings

The backbone has 24 double-stream blocks, hidden width 3,072, 24 attention heads, 48 latent channels, and approximately 8.49B parameters. The temporal and spatial RoPE dimensions are $[32, 48, 48]$. The

conditioning projection expands 48 channels to 97, adding $49 \times 3072 = 150{,}528$ weights. Reference slots retain unit loss weight. Full bidirectional attention uses text-padding masking. Reference conditioning uses channel concatenation, and training uses the reconstruction objective. Table 8 summarizes the training and inference settings.

Table 8: Main fine-tuning and inference configuration. Conditioning probabilities act at the segment/anchor level; the resulting task frequencies depend on reference availability and selection.

Setting	Value
Pure T2V, steps 0–19k	19,000 updates, 256 H800 GPUs, LR 10^{-5}
Reference fine-tuning, steps 19k–36k	17,000 updates, 16 H800 GPUs, LR 10^{-5}
Reference fine-tuning, steps 36k–45k	9,000 updates, 16 H800 GPUs, LR 2×10^{-6}
Global batch size, steps 0–19k	128 videos per optimizer update
Global batch size, steps 19k–36k	8 videos per optimizer update
Global batch size, steps 36k–45k	8 videos per optimizer update
Training clip length	241 frames
Spatial preset	480p, aspect-ratio dependent
Optimizer	AdamW
Warmup, continuation	200 updates
Configuration seed	42
Per-segment condition probability	0.6
Image-anchor probability	0.8
Condition dropout probability	0.2
Conditioning path	Sparse reference placement, channel concatenation
Shot-coordinate offset	$\alpha = 0.5$
Sequence-parallel size, final fine-tuning	8
Inference steps / CFG / flow shift	50 / 6.0 / 7.0

Training data and task sampling. The first reference-training stage fits references to their target lengths for 17,000 steps. Its base task sampler assigns probabilities 0.3 to text-only generation and 0.7 to reference conditioning. Dataset references are sampled separately, with probability 0.6 per available video reference and 0.8 for an image anchor. Selected dataset references activate conditioning regardless of the base task draw. Condition dropout removes individual references with probability 0.2 while retaining at least one when references are active.

The final 9,000 updates use all 297,192 source-video records with the filtered DurMix-300K conditions and retain 117,868 records from the earlier reference data, a 71.60:28.40 ratio by record count. References retain their native lengths and are placed sparsely on the target timeline. Dataset reference sampling remains active, while examples without selected references use text-only training. Both the main results and the complete coordinate-by-placement comparison use 45,000 training steps.

C Evaluation Sets and Metrics

The main benchmark evaluates unified generation on 300 requests: 30 T2V, 30 TI2V, 120 V2V, and 120 TIV2V.

In each of V2V and TIV2V, 40 requests (one third) replace the original descriptions of video-referenced shots with reference-control instructions. The remaining 80 requests retain those descriptions and append the control instructions. The shared style prompt and descriptions of shots without video references are retained, including image-anchor shots. The instructions are:

- **Video reference:** "Use the provided reference video segment for this time range and preserve its subject identity, motion, camera movement, and temporal dynamics."
- **Image reference:** "Use the provided reference image anchor at latent frame t and preserve the subject identity and appearance around that moment."

Here t specifies the image-anchor position; image instructions are appended to the corresponding shot descriptions. During evaluation, image anchors are processed before video references. A reference whose destination interval overlaps an already accepted reference is skipped with a warning.

Transition timing. We match TransNetV2-detected cuts to the requested boundaries using order-preserving dynamic programming. Matched cuts contribute absolute timing error, and each missing or extra cut adds a one-second penalty. The total timing error in seconds is normalized by the larger cut count. Transition MAE averages the normalized timing error over videos.

Text alignment. CLIP-T averages CLIP (Radford et al., 2021) similarity between generated frames and their corresponding shot descriptions. TA scores each requested shot with the pretrained VideoAlign model and its published normalization (Liu et al., 2025a), yielding a signed reward. TA uses the original shot description and shared style prompt, including for shots whose inference descriptions were replaced by reference-control instructions. Scores are averaged across shots and then across videos.

Content consistency. We adapt the consistency categories of MultiShotMaster (Wang et al., 2026b) to requests that explicitly change subjects and scenes. For each request, a fixed annotation of the original shot descriptions identifies recurring entities, shared scenes, and semantically comparable shot pairs. These annotations are shared by all methods. Inter-Shot Semantic averages ViCLIP (Wang et al., 2024) cosine similarity over the annotated pairs, using eight uniformly sampled frames per requested shot. Inter-Shot Subject compares only the same entity across shots, and Inter-Shot Scene compares visible backgrounds only for shots annotated as sharing a scene.

GroundingDINO (Liu et al., 2024) and SAM (Kirillov et al., 2023) locate and segment subjects. We compute DINOv2-small cosine similarities on subject crops with non-subject pixels masked, or on full frames with foreground regions masked for background comparisons. First, middle, and last frames are sampled independently from each interval; similarities average all available keyframe pairs. Reference Subject and Reference Background use the corresponding reference and destination shot. Images use their exact anchor frames. Deliberate subject-appearance or material edits are excluded from reference-subject comparisons; background, layout, or material edits are excluded from reference-background comparisons. Reference Background excludes 82 references with background, spatial-layout, or material/color edits.

Reference Semantic measures ViCLIP cosine similarity between each video reference and its corresponding generated shot. Eight frames are uniformly sampled from the full native reference and, independently, from its requested destination interval. Each eight-frame sequence is resized to 224×224, normalized, and encoded by ViCLIP. Cosine similarities are averaged over references within each video and then across videos. The evaluation covers all 489 video references in 240 V2V/TIV2V requests, including all edit categories. Image anchors are excluded from this video-level metric. Destination intervals follow the fixed benchmark timeline used for the other content metrics, without adjustment to detected cuts or output length. Each reported method has valid scores for all 489 pairs.

Each metric compares methods on the same shot or reference pairs and weights videos equally. These feature similarities measure cross-shot content correspondence. SDC uses detected shots as defined in Section 3.3.

C.1 Baseline Configurations

Generation protocol. Phantom, RefAlign, StoryMem, and VACE generate individual shots and concatenate them at the requested boundaries. Each shot receives the shared style description and its own conditioning text, and all shots within a request use the same random seed. Concatenation uses hard cuts without overlapping frames or transition blending. For models requiring a $4k + 1$ frame count, the extra terminal frame of each non-final shot is removed before concatenation. Since this procedure sets the shot boundaries externally, Transition MAE is unreported for these four methods. HunyuanVideo1.5, MiniMax-H3, and Refrain generate the complete sequence jointly and are evaluated for transition timing.

HunyuanVideo1.5. We evaluate the pretrained backbone before our multi-shot and reference-conditioned training. Shot-aware temporal offsets are disabled. Image and video references are sparsely placed in the existing conditioning channels, together with the occupancy mask. The output latents remain generated throughout denoising, including at reference locations. Sampling uses 50 steps, guidance scale 6.0, flow shift 7.0, and sequence-parallel size 8.

Phantom. We use the released Phantom-14B subject-conditioned model for shots with references. Each video reference is represented by its first native frame, and image references are assigned to their corresponding shots. This adaptation supplies appearance information through the model's image-reference interface. Reference-conditioned shots use 40 UniPC steps with text guidance 7.5, image guidance 5.0, and flow shift 5.0. Text-only shot generation uses 50 steps with guidance 5.0 and flow shift 5.0.

RefAlign. We use the released RefAlign-14B checkpoint with its DINOv3 reference encoder. Each video reference contributes its first native frame, while image references are supplied directly. References are routed to the corresponding shot through the subject-image interface. Image anchors therefore condition their owning shots through this interface. Each shot is generated with 50 steps, text guidance 5.0, reference guidance 7.5, and flow shift 5.0.

StoryMem. We use the released memory-conditioned model and its visual-memory mechanism. Each video reference contributes its first, middle, and last native frames, with duplicate positions removed for short clips. Image references are added directly to the initial memory. These images form global visual memory available across shots. After each shot, generated keyframes are added to the memory, retaining the initial references and the most recent generated frames within a ten-image limit. Memory-conditioned generation uses 40 UniPC steps, guidance 3.5, and flow shift 4.0. For text-only requests, Wan2.2-T2V-A14B generates the first shot, and subsequent shots use generated memory. Native 16-fps outputs are mapped to the requested 24-fps timeline by nearest-frame selection before concatenation.

VACE. We evaluate Wan2.1-VACE-14B on the image- and video-conditioned requests using sparse reference placement. Native reference frames are placed at selected positions in each shot's input canvas without temporal interpolation. Image anchors occupy their assigned positions. Occupied pixels are marked as known regions, and the remaining positions are marked for generation. Sampling uses 50 UniPC steps, guidance 5.0, and flow shift 16.0. The internal canvas respects the implementation's 480p area and sequence-length limits, and generated shots are spatially fitted to the requested output geometry before concatenation. Reference Semantic, Subject, and Background are left unreported for this completion-based adaptation, in which supplied reference regions are reconstruction conditions.

MiniMax-H3. We run the model locally on NVIDIA H800 GPUs through its Diffusers implementation, using one inference process per GPU. Text-only requests use the text-to-video-and-audio checkpoint, while image- and video-conditioned requests use the reference-conditioned checkpoint. Images and video clips are passed through the native reference interfaces and identified in the prompt by their corresponding picture and video tags. Generation uses bfloat16 weights, 50 denoising steps, and the released guidance-distilled sampling configuration. The model jointly generates video and audio, and evaluation uses the video output.

Prompts specify shot indices and explicit cut times in seconds. Output lengths are aligned to the model's temporal grid, and cut timestamps are scaled to the submitted duration. Transition MAE is measured against these submitted timestamps. Spatial dimensions are rounded up to the required multiple of 32 and center-cropped to the requested size after generation. Reference clips shorter than the implementation's 13-frame minimum are expanded to 13 frames by nearest-frame sampling with their endpoints preserved. Prompt availability follows the protocol above: 40 requests in each of V2V and TIV2V omit the original descriptions of video-referenced shots, while the remaining requests retain them.

Metric reporting and computational cost. Crosses in Table 1 indicate unreported or inapplicable measurements. Transition MAE is omitted for methods whose boundaries are fixed by shot concatenation. Reference Semantic, Subject, and Background are unreported for VACE and the pretrained HunyuanVideo1.5 baseline. All reported SDC scores use the same real-video calibration bank.

FLOPs are analytical denoising estimates for a 480×832, 241-frame video with three shots and three 49-frame references. The estimates include each method's sampling schedule, guidance branches, reference processing within the denoiser, and active adapters. MiniMax-H3's joint audio-token computation is included. Text and visual encoders, VAE processing, data transfer, and distributed

communication are excluded consistently. Sequence parallelism partitions the computation without reducing its total FLOPs.

D SDC Details and Validation

Sampling and representations. Both generated and real videos use TransNetV2 shot boundaries. We remove three frames at each boundary and sample two to four distinct, uniformly spaced interior frames per shot. SDC requires at least two usable shots. Content uses DINOv2-small CLS embeddings; style uses the CSD ViT-L/14 style head and VGG16 (Simonyan & Zisserman, 2015) channel means and standard deviations from `relu1_2`, `relu2_2`, `relu3_3`, and `relu4_3`. For VGG, each layer's concatenated moments are normalized before concatenation across layers. All final frame embeddings are L_2-normalized.

Energy contrast and closed form. For frame features $x^r_{i,a}$, define $(D^r_{ij})_{ab} = 1 - \langle x^r_{i,a}, x^r_{j,b}\rangle$ and the shot mean $\mu^r_i = n_i^{-1}\sum_a x^r_{i,a}$. Distance averages include every entry, including the zero within-shot diagonals. Thus,

$$\begin{aligned} E^r_{ij} &= 2\,\mathrm{mean}(D^r_{ij}) - \mathrm{mean}(D^r_{ii}) - \mathrm{mean}(D^r_{jj}) \\ &= 2(1 - \langle \mu^r_i, \mu^r_j\rangle) - (1 - \|\mu^r_i\|^2) - (1 - \|\mu^r_j\|^2) \\ &= \|\mu^r_i - \mu^r_j\|^2. \end{aligned} \tag{12}$$

Subtracting the within-shot distance terms gives a comparison of sample means, equivalent to empirical squared MMD with a linear kernel (Székely & Rizzo, 2013; Gretton et al., 2012; Sejdinovic et al., 2013). Floating-point round-off is handled by setting energy values in $[-8\epsilon_{32}, 0)$ to zero, where ϵ_{32} denotes float32 machine precision. Values below $-8\epsilon_{32}$ are rejected as invalid.

Within-shot calibration. To measure natural variation without introducing a shot transition, we partition each shot's sampled frames into balanced, disjoint groups and compute their content and style energies. Each unordered split is used once: two, three, and four frames yield one, three, and three splits, respectively. The bank contains 2,144,201 real videos, 6,462,692 shots, and 19,388,074 splits, with all three representations available on the same frames. All but six shots have four sampled frames.

For four sampled frames, define $d_a = x^r_{i,a} - \mu^r_i$, which satisfies $\sum_a d_a = 0$. The three partitions Π_i pair frame 1 with $j \in \{2,3,4\}$, giving

$$\begin{aligned} \frac{1}{3}\sum_{\pi\in\Pi_i} E^r_{i,\pi} &= \frac{1}{3}\sum_{j=2}^{4} \|d_1 + d_j\|^2 \\ &= \frac{1}{3}\sum_{a=1}^{4} \|d_a\|^2 = \mathrm{tr}(\widehat{\Sigma}^r_i). \end{aligned} \tag{13}$$

Here $\widehat{\Sigma}^r_i$ is the sample covariance with divisor three: the average split energy measures within-shot feature dispersion. For n frames split into groups of sizes k and $n-k$, the average is $n\,\mathrm{tr}(\widehat{\Sigma})/(k(n-k))$, with covariance divisor $n-1$. The same factor applies to content and style features.

Calibration groups and source exclusion. The real-video bank is partitioned by source video into three disjoint calibration subsets, indexed by $g \in \{1,2,3\}$. Each subset provides a separately fitted content–style slope and conditional residual distribution. Averaging their ranks combines the three calibration estimates into a style-head rank (Section 3.3).

Source exclusion uses a separate five-fold partition to prevent a query's source video from contributing to its calibration. After shuffling unique source videos with seed 20260908, zero-based index i determines group $g = 1 + (i \bmod 3)$ and fold $i \bmod 5$. The query source's fold is excluded from both slope fitting and residual ranking in every group. Outputs from all methods and edited variants of the same source use the same excluded fold. Sources absent from the bank receive a fixed fold based on their identifier. The two style heads, three groups, and five excluded-fold configurations require 30 calibration fits, all using real videos.

Robust content correction. For each group g and style head r, we fit style energy s_b against content energy c_b using Huber regression (Huber, 1973). A bank split has weight $w_b = 1/(N_g S_v B_{v,q})$, where N_g counts the videos remaining in group g after source-fold exclusion, S_v counts usable shots in video v, and $B_{v,q}$ counts splits in shot q. These weights assign equal mass to each video, distributed equally among its shots and then their splits. Suppressing group, head, and excluded-fold indices, we initialize κ with the weighted median of s_b/c_b for $c_b > \epsilon_{32}$ and iterate

$$\begin{aligned} z_b &= s_b - \kappa c_b, \qquad m = \mathrm{wmed}_w(z_b), \\ \sigma &= 1.4826\ \mathrm{wmed}_w(|z_b - m|), \\ a_b &= \min\left(1, \frac{1.345\sigma}{\max(|z_b|, \epsilon_{64})}\right), \\ \kappa_{\mathrm{new}} &= \frac{\sum_b w_b a_b c_b s_b}{\sum_b w_b a_b c_b^2}. \end{aligned} \tag{14}$$

The weighted median m centers residuals for the scale estimate σ. The through-origin fit uses uncentred residuals z_b in its influence weights. Here ϵ_{64} is float64 machine precision. Iteration stops at $|\kappa_{\mathrm{new}} - \kappa| \le 10^{-12}(1 + |\kappa|)$, after 100 iterations, or when $\sigma \le \epsilon_{64}$. Query residuals $e^r_{ij,g} = E^r_{ij} - \kappa_{g,r} E^c_{ij}$ retain their sign. The residual correction subtracts the fitted linear content–style trend. At fixed content energy, subtracting the same linear term preserves style ordering. Queries outside the range of bin centers in content-percentile space use the nearest bin.

Responses by edit type. In the single-shot style-edit comparison in Table 4(a), one shot receives a style-transfer edit and the other remains unedited, with the corresponding original shot pair serving as the control. The shared-versus-independent comparison in Table 4(b) uses the same two reference-shot positions in both conditions, with both shots edited. Comparisons retain matched shot pairs with valid scores for all compared metrics. Scores are averaged over eligible pairs within each video and then equally across videos. For each metric, the fifth percentile of the original-video scores in (a) or the shared-edit scores in (b) defines the detection threshold. The reported detection rate is the fraction of edited videos scoring below this threshold.

Table 9 separates the responses by edit type. SDC decreases more for style, filter, and lighting edits than for motion or spatial changes.

Table 9: **Sensitivity to reference edit types.** Relative decrease in mean similarity or consistency from original shot pairs to pairs with one edited shot. Larger decreases indicate stronger responses within each metric's scale. A source video can appear in multiple edit categories.

Edit type	Videos	SDC	CLIP	ViCLIP	DINO
Camera / lens filter	21,699	96.2%	13.4%	33.9%	22.0%
Style transfer	108,909	86.5%	14.9%	30.1%	22.8%
Lighting / weather	29,516	75.0%	10.4%	22.5%	17.1%
Background retexture	27,253	64.2%	10.2%	24.4%	21.0%
Object material / color	23,405	51.0%	7.8%	16.6%	10.1%
Subject appearance	29,873	32.4%	9.5%	18.0%	13.5%
Motion change	22,559	19.2%	4.6%	7.1%	3.8%
Spatial relation change	20,424	18.6%	4.0%	6.3%	3.2%

Comparison with style-feature similarity. Table 10 compares SDC with CSD and VGG cosine similarity using the same frame features and video-level averaging. Each cosine score averages all cross-shot frame pairs of normalized style features. The edit comparisons follow the control thresholds defined above. For the real-video comparison, each metric's fifth percentile over all 2,144,201 real videos defines a low-score threshold. The last column reports the fraction of the 458,920 videos with DINO similarity below 0.25 falling below that threshold. CSD achieves the highest detection rates in both edit comparisons. SDC retains sensitivity to style conflicts while reducing score dependence on content changes, with a low-score rate of 4.4% on the low-DINO subset compared with 18.8% for CSD and 13.7% for VGG.

Effect of linear correction. We set $\kappa = 0$ for both bank and query energies in both style heads, keeping the features, calibration videos, source-fold exclusion, bin boundaries, sample weights,

Table 10: **Style-feature similarity and SDC.** Edit detection uses a 5% control false-positive rate. Low-score rates use each metric's fifth percentile over all real videos as the threshold.

Metric	Single-shot style edits Detection ↑	Independent vs. shared edits Detection ↑	Real videos, DINO < 0.25 Low-score rate
CSD cosine	78.6%	58.5%	18.8%
VGG cosine	34.9%	39.5%	13.7%
SDC	78.0%	50.9%	4.4%

interpolation, and fusion fixed. Within each bin, bank samples are reordered by uncorrected style energy. The comparison reuses the 108,909 single-shot style-edit videos, 72,570 shared/independent-edit videos, and 2,144,201 real videos above. Thresholds follow the preceding comparison and are recalculated for each variant.

Table 11: **Effect of linear correction with fixed conditional ranking.** Edit detection uses a 5% control false-positive rate. The last column gives the low-score fraction among 458,920 real videos with DINO < 0.25, using each variant's fifth percentile over all real videos as its threshold.

Setting	Single-shot style edits Detection ↑	Independent vs. shared edits Detection ↑	Real videos, DINO < 0.25 Low-score rate
$\kappa = 0$	76.79%	50.29%	6.01%
Full SDC	77.98%	50.85%	4.38%

Linear correction increases the two edit-detection rates by 1.19 and 0.56 percentage points (Table 11). Their 95% intervals are [0.95, 1.42] and [0.23, 0.82] percentage points, respectively. The low-DINO real-video low-score rate decreases by 1.63 points, with a 95% interval of [1.60, 1.67]. Intervals use 2,000 paired source-video bootstrap draws, re-estimating each variant's threshold in every draw while holding the calibration bank fixed.

For real-video shot pairs, the mean absolute score change is 0.0060 when ranks interpolate between bin centers and 0.0728 when a query uses an endpoint bin in at least one calibration group or style head. Across all real videos, Spearman correlation with DINO is -0.0879 with correction and $+0.0184$ at $\kappa = 0$. Thus, linear correction gives modest detection gains, while $\kappa = 0$ has a global content correlation closer to zero. The largest score changes occur for queries using endpoint bins.

D.1 Calibration Stability

Table 12 evaluates six sample sizes from 100 to all 2,144,201 usable real videos. We refit the CSD/DINOv2 slope on each subset using Equation 14 and the bank's video–shot–split weights. Each size has five disjoint subsets up to 100,000 videos, two at one million, and one at the full size. Subsets at different sizes can overlap. SDC is averaged over these same query subsets using the fixed full real-video bank with source-fold exclusion.

Mean κ ranges from 0.39100 to 0.40168, reaching 0.39874 on the full pool. Its standard deviation across repeated fits decreases from 0.01533 at 100 videos to 0.00014 at one million. Mean SDC ranges from 0.33636 to 0.34414 and is 0.33731 on the full pool. The two columns measure slope-estimation stability and the sampling stability of mean SDC under a fixed bank, respectively.

With the source population, encoders, and sampling protocol fixed, larger samples estimate the same content–style relationship more precisely. These results support using the real-video bank as a stable reference for comparison. All generators use this shared bank for both slope correction and conditional ranking.

Table 12: **Real-video κ and SDC across six sample sizes.** Repeated subsets are disjoint within each size. We refit κ on each subset and average SDC over the same videos using the fixed full real-video bank. Standard deviations and ranges summarize repetitions; the last row uses all videos.

		κ				SDC	
Videos	Repeats	Mean	sd	Minimum	Maximum	Mean	sd
100	5	0.39100	0.01533	0.36996	0.41107	0.33636	0.02372
1,000	5	0.40168	0.00284	0.39684	0.40415	0.34414	0.00484
10,000	5	0.39920	0.00065	0.39813	0.39984	0.33795	0.00227
100,000	5	0.39890	0.00066	0.39793	0.39982	0.33711	0.00071
1,000,000	2	0.39872	0.00014	0.39858	0.39886	0.33733	0.00002
2,144,201	1	0.39874	—	0.39874	0.39874	0.33731	—

E HUMAN STUDY

Human judgments. Six professional raters independently assessed 120 clips in randomized order, with method identities, SDC scores, and edit assignments hidden. The set contains 80 Refrain-generated clips sampled across SDC quintiles before annotation, 20 real clips, and 20 spliced controls. Each control joins two edited shots from the same source video. Ten controls share an editing instruction and ten use different edit categories. Raters assessed colour, lighting, contrast, and texture on a five-level scale: 1 indicates clearly inconsistent style, 3 is borderline, and 5 indicates clearly consistent style. Instructions distinguish style conflicts from changes in subject or viewpoint. All six raters assessed every clip, providing 720 style judgments.

Consistency labels. We analyze the five-level annotations as a binary discrimination task. Ratings of 1–2 indicate inconsistency, ratings of 4–5 indicate consistency, and 3 is undecided. A clip receives a class label when at least four of six raters agree. This yields 104 consistent clips and eight inconsistent clips. The remaining eight clips have no majority class and are excluded from AUC, with their ratings and SDC scores retained in the analysis records. Labels follow human judgments for all sources, including the spliced controls.

SDC scoring and discrimination. All clips are scored using the same real-video bank and CSD/VGG fusion as the main evaluation, with source-fold exclusion. For the 20 spliced controls, features are extracted directly from the videos shown to raters, using their recorded splice boundaries and the same frame-sampling procedure as Appendix D. AUC measures the probability that a human-consistent clip receives a higher SDC score than a human-inconsistent clip, with half credit for ties. Table 13 shows the score separation: mean SDC is 0.3161 for consistent clips and 0.0360 for inconsistent clips.

Table 13: **SDC by human consistency judgment.** Class labels require at least four of six votes. Eight unresolved clips are excluded from AUC and retained in the annotation records.

Human judgment	Videos	Mean SDC	Median SDC
Consistent	104	0.3161	0.2872
Inconsistent	8	0.0360	0.0195

Uncertainty. The AUC is 0.969, with a 95% interval of [0.931, 0.995] from 20,000 source-video bootstrap draws. Five draws containing only one class are omitted because AUC is undefined for them. A one-sided test uses 100,000 permutations of human labels across source blocks of equal size, preserving source dependence and class counts. No permutation reaches the observed AUC. The plus-one correction gives $p = 1/(100{,}000 + 1) \approx 10^{-5}$. These results assess consistency discrimination on the annotated set, which contains eight consensus-inconsistent clips.

F LIMITATIONS

Refrain is primarily evaluated on short multi-shot sequences, and maintaining recurring identities and a coherent style across substantially longer narratives requires further evaluation. The current task uses a shared style prompt to specify the desired visual appearance across shots, while deliberate shot-specific style changes require a more detailed description of the intended output. Explicit control over which reference attributes to preserve or revise remains a direction for future work. DurMix-300K pairs edited references with their original videos, so its coverage is bounded by the source footage and editing transformations used in construction. SDC measures cross-shot style consistency relative to a fixed feature space and calibration bank, and its scores should be interpreted alongside reference fidelity, motion quality, and human judgments.

G FUTURE WORK

A central direction is consistency-guided post-training, using SDC together with human preferences to improve cross-shot style consistency while preserving reference content and motion. Comparing candidate videos generated from the same prompts and references could provide preference supervision, with reference fidelity and motion quality included in candidate selection. We plan to examine whether this supervision improves consistency across all four conditioning tasks and how well these gains agree with human judgments. We also aim to extend Refrain to longer narratives and references collected from independent videos, testing recurring subjects and locations across longer shot sequences and a broader range of visual domains. Finally, we plan to explore localized revision, allowing users to replace a reference or revise an individual shot while preserving the timing, content, and style of the remaining sequence.

H ADDITIONAL QUALITATIVE RESULTS

H.1 REFERENCE DURATION AND MOTION PACING

Variable-duration reference control is a core capability of Refrain: the requested shot can be longer or shorter than its reference, while the selected reference start and target interval remain unchanged. Uniformly stretching or compressing the reference changes motion speed, while truncating it can discard important actions. The examples below examine both directions, showing how reference actions are adapted while preserving natural pacing and a shared style across shots.

A shorter reference. Figure 6 (`case12_v2v` in the supplementary videos) uses a 3.7-second reference for a 6.5-second target shot. Both Refrain and MiniMax-H3 maintain broadly similar action speeds to the reference, preserving the pace of the camera gesture. Their cross-shot appearance differs: Refrain brings the referenced shot into the surrounding shots' visual style, whereas MiniMax-H3 retains the reference's warm orange background and produces a visible style mismatch across shots.



Figure 6: **Short reference, longer target: 3.7s → 6.5s.** Both methods preserve plausible action pacing; Refrain also maintains a shared style across shots. Dashed arrows link reference and generated action states.

A longer reference. Figure 7 (`case13_v2v` in the supplementary videos) uses a 3.9-second reference for a 2.8-second target shot. Refrain condenses the action sequence while retaining the woman's reaction, hand gesture, and subsequent head turn at a natural pace. MiniMax-H3 omits more of this progression, keeping the woman in a similar side-facing pose through much of the shot. The methods differ in the actions retained within the shorter target interval.



Figure 7: **Long reference, shorter target: 3.9s → 2.8s.** Refrain condenses the reference action while retaining its key gestures and natural pacing. MiniMax-H3 omits more intermediate actions.

H.2 Further Generation Examples

Style: *Photorealistic documentary footage, unfinished construction interior, natural lighting, subtle handheld camera movement.*
Shot 1 [0.0s-1.0s]: *Low-angle close-up reveals an old electrical junction surrounded by tangled wires.*
Shot 2 [1.0s-2.3s]: *The camera slowly pushes closer, revealing screws, connectors, and tangled wiring.*
Shot 3 [2.3s-3.6s]: *A man's hands twist exposed copper wires beneath unfinished ceiling beams.*
Shot 4 [3.6s-5.0s]: *His finger points inside the junction, inspects it, then withdraws.*



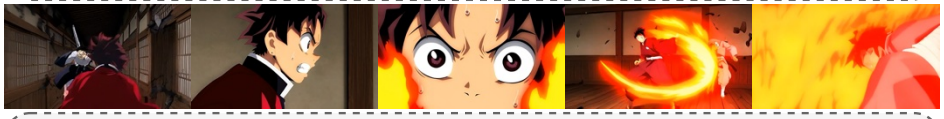
Style: *Japanese action anime, bold linework, exaggerated movement, dramatic shadows, glowing blue energy illuminating a barren canyon.*
Shot 1 [0.0s-0.3s]: *Close-up: a wounded, black-haired man screams furiously, his brows tightly furrowed.*
Shot 2 [0.3s-0.9s]: *He braces against a green robot's blue beam; yellow rings expand.*
Shot 3 [0.9s-2.4s]: *Side close-up reveals clenched teeth, strained muscles, and a scratched arm.*
Shot 4 [2.4s-3.4s]: *The man continues to block the green robot's blue energy beam with his body.*
Shot 5 [3.4s-3.8s]: *Extreme close-up: the green robot's enormous eye stares coldly forward.*
Shot 6 [3.8s-5.0s]: *Through a targeting interface, his strained face remains painfully determined.*



Style: *Realistic cinematic drama with bright natural sunlight, an urban sidewalk setting, shallow depth of field, and tense, emotionally charged performances.*
Shot 1 [0.0s-2.2s]: *Focus shifts from the driver's mirror reflection to two people exiting.*
Shot 2 [2.2s-4.2s]: *Close-up: the driver stares at his smartphone, speaking in disbelief.*
Shot 3 [4.2s-5.0s]: *An older man holds the woman's waist, then kisses her cheek.*



Style: *Japanese battle anime, grounded action choreography, a ruined traditional wooden corridor at night, intense orange-red fire effects.*
Shot 1 [0.0s-1.5s]: *Amid wreckage, a young swordsman's red haori becomes green-and-black checkered.*
Shot 2 [1.5s-2.1s]: *Side close-up: the sweating swordsman clenches his teeth, glaring determinedly.*
Shot 3 [2.1s-3.1s]: *The camera rushes toward his furious face as flames emerge.*
Shot 4 [3.1s-3.9s]: *He unleashes a blazing sword arc toward his bracing opponent.*
Shot 5 [3.9s-5.0s]: *Extreme close-up: flames strike the enemy's shoulder, violently twisting his body.*



Style: *Photorealistic contemporary urban drama, modern interiors, bright diffused lighting, minimal shadows, clean and slightly cool visuals.*
Shot 1 [0.0s-1.5s]: *A ponytailed woman enters in a white T-shirt, visibly astonished.*
Shot 2 [1.5s-2.2s]: *Prepared dishes cover a black-and-gold marble table; nobody is present.*
Shot 3 [2.2s-3.1s]: *Close-up: her shock fades into resolve, lips pressed, brows furrowed.*
Shot 4 [3.1s-5.0s]: *She stands determined, turns left, and exits beside the staircase.*



Figure 8: **Text-only multi-shot generation.** Examples span documentary, animation, and drama. Each request specifies a shared style and per-shot events.



Figure 9: **Reference-conditioned generation (1/2).** Each example shows independently edited references, the requested shared style, and generated shots.



Figure 10: **Reference-conditioned generation (2/2).** Further examples under the protocol of Figure 9.