SEGUE: VISUAL SIGNAL AS EXCLUSIVE GUIDANCE FOR ONE-SHOT TRANSITION EFFECT GENERATION

Anonymous authors
Paper under double-blind review

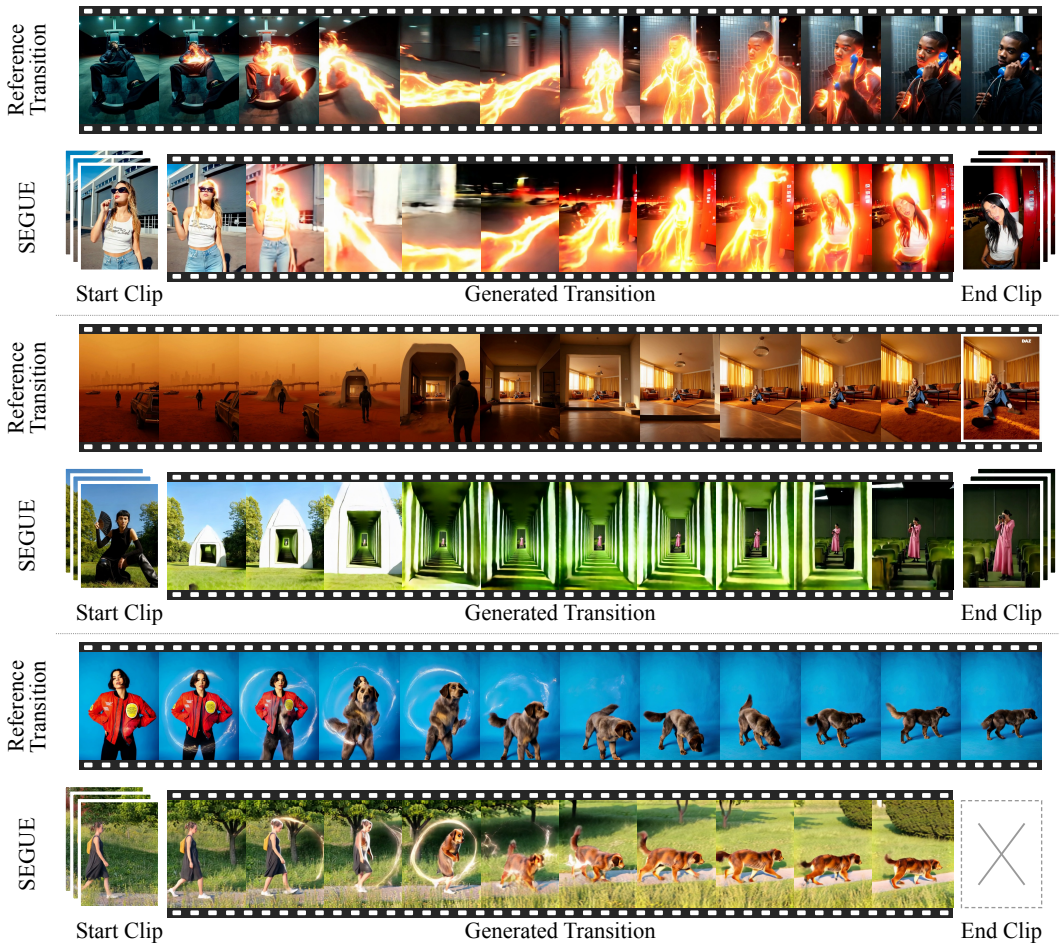

Figure 1: **SEGUE** is a one-shot transition effect generator: it successfully connects two endpoints (a video clip, a static image, or absent) given a reference transition. Videos available in supplemental.

ABSTRACT

One-Shot Transition Effect Generation (TEG) aims to bridge two video clips using a transition effect provided in a reference clip. However, existing methods learn a text shortcut: they rely heavily on prompts with detailed effect descriptions, and the visual transition signal loses its weight. We present SEGUE, a framework designed to read the transition strictly from the visual demonstration via two key mechanisms: i) **Visual reference isolation** during training, pairing inputs with effect-free neutral prompts and applying a one-way attention mask that keeps the reference independent of the target; and ii) **Null-reference guidance** at inference, which defines the unconditional path as a dissolve of the reference endpoints—adding no motion or material of its own—to isolate and amplify the visual transition effect. SEGUE flexibly supports endpoints that are video clips, static images, or absent, unifying two-sided TEG and single-sided visual effect transfer within a single model. We also introduce the SEGUE-Benchmark, a cross-paired dataset of 1,692 transition classes and 56,368 training pairs spanning creative presets, parametric

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.

procedural transitions, and visual effects. On held-out effects, SEGUE outperforms baselines adapted to TEG and previous visual effect transfer methods even without a detailed effect description in the prompt, serving as the first open-source framework for one-shot TEG.

1 INTRODUCTION

Transitions shape the pacing and continuity of a video, deciding how one shot gives way to the next (Zhang et al., 1993; Soucek & Lokoc, 2024; Shen et al., 2022). Beyond simple cuts or dissolves, a *transition effect* performs this hand-off with its own visual dynamics, such as volumetric smoke, opening portals, or wipes, and brings in motion and material that belong to neither endpoint (Liu et al., 2025; Mao et al., 2026). While video diffusion models can bridge clips (Chen et al., 2023; Yang et al., 2025; Kan et al., 2026; HaCohen et al., 2026), transferring an arbitrary transition effect from a reference clip remains challenging. The model must connect unrelated scenes and transfer the effect's dynamics without leaking the reference's own scenes into the output (Li et al., 2025a; Jones et al., 2026; Yin et al., 2026). We formalize this task as One-Shot Transition Effect Generation (TEG, Figure 1)[1]: given a reference clip showing a transition effect and a target pair of start and end clips, synthesize that effect between them.

Current reference-conditioned approaches are primarily developed for visual effect transfer onto a single starting frame or a full video (Li et al., 2025a; Bian et al., 2026; Jones et al., 2026; Yang et al., 2026a). They pair the visual reference with a text prompt that describes the effect in detail, which we call the effect prompt. This creates a **text shortcut**: because text provides a compact semantic signal, models learn to infer the effect from the prompt, leaving the visual reference under-utilized (Figure 2). As a result, once the effect description is removed from the prompt, the visual reference signal cannot guide the generation and the transition quality drops. More importantly, text cannot convey the exact motion and timing of an effect, and has no words for a novel, uncataloged effect. Because the true dynamics exist only in the video frames, the model must read and transfer the transition directly from the visual demonstration.

To address this limitation, we introduce SEGUE, a framework that reads the transition strictly from the visual demonstration, through visual reference isolation during training and null-reference guidance at inference. During training, SEGUE enforces reference isolation through two complementary designs: neutral text conditioning that contains no description of the intended effect, to close the text shortcut, and a one-way attention mask that allows the target to read the reference while preventing reference representations from being rewritten. At inference time, SEGUE amplifies this visual signal using *null-reference guidance* (NRG). Rather than contrasting against an empty condition, which removes both the effect and the reference scenes and so amplifies both, NRG contrasts against a null reference that keeps the scenes of the reference and removes only the transition effect, built as a dissolve of the reference's own endpoints. Because this dissolve adds no motion or material of its own, steering away from it amplifies the transition dynamics and cancels out the leakage of the reference scenes. Finally, SEGUE supports endpoints that are video clips, static images, or absent, unifying two-sided TEG and single-sided visual effect transfer within a single model.

Because no existing benchmark supports training or evaluating reference-conditioned transitions, we introduce the **SEGUE-Benchmark**. It establishes a comprehensive dataset spanning creative preset transitions, parametric procedural transitions, such as geometric wipes, volumetric zooms, and fluid warps, and complex physical visual effects. On held-out effects, SEGUE outperforms baselines adapted to TEG and surpasses previous visual effect transfer methods, even though both rely on text prompts containing the effect description, while SEGUE operates on the isolated visual reference alone.

In summary, our key contributions are:

- We formalize One-Shot Transition Effect Generation (TEG) and introduce SEGUE, the first model designed for one-shot transition effect generation, capable of flexibly transferring transitions across arbitrary endpoints, whether each is a video clip, a static image, or absent.
- We propose a two-part method to isolate and amplify the visual reference signal: training-side *visual reference isolation*, which combines a neutral text prompt with a one-way attention mask to

[1] A transition effect goes from the start clip to the end clip, whereas a visual effect is conditioned on the start endpoint alone and ends freely, as in the last row of Figure 1. We call the first task TEG and the second visual effect transfer.

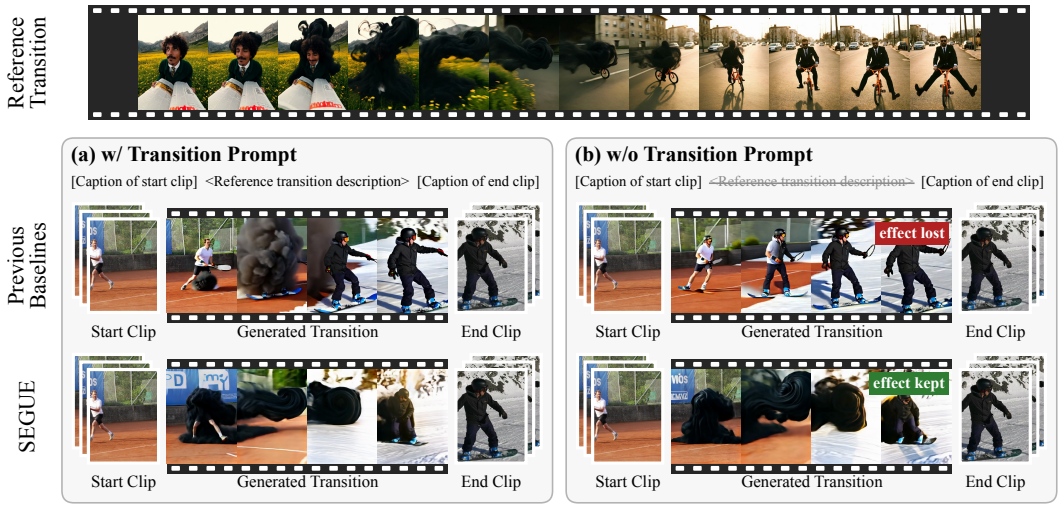

Figure 2: **The text shortcut.** Given the same visual reference (top) and endpoints, previous methods rely heavily on the effect prompt and under-utilize the visual reference; with the neutral prompt (b), they fail to produce the transition and degrade in quality. In contrast, SEGUE reads the dynamics directly from the visual reference, preserving the effect with the effect prompt (a) and with the neutral prompt (b).

close the text shortcut and isolate the visual reference, and inference-side *null-reference guidance*, which contrasts against a dissolve of the reference endpoints to amplify the transition dynamics while suppressing the leakage of the reference scenes.

- We build the SEGUE-Benchmark: a cross-paired dataset of 1,692 transition classes and 56,368 training pairs, spanning creative presets, parametric procedural transitions, and visual effects, together with an evaluation dataset and the *transition fidelity* metric.

2 Related Work

Controllable video generation. Pretrained video generators are steered by structure and motion through control adapters (Zhang et al., 2023; Guo et al., 2024b; Wang et al., 2024b), and by example videos that carry motion (Zhao et al., 2024b; Wei et al., 2024; Xu et al., 2026; Ressler-Antal et al., 2025), actions (Sun et al., 2024), physics (Chen et al., 2026), or edits (Wang et al., 2026). Reading a task from one visual example goes back to image analogies (Hertzmann et al., 2023) and visual prompting (Bar et al., 2022; Wang et al., 2023a;b), and image models now do it with little or no text (Gu et al., 2024; Li et al., 2025c; Chen et al., 2025). In-context conditioning places the example in the same token sequence as the target (Huang et al., 2024a; Fei et al., 2024), so a clip, an image, or nothing can fill an input slot without a dedicated branch (Jiang et al., 2025). Guidance amplifies a condition by contrast with an empty condition (Ho & Salimans, 2022), a weaker model (Karras et al., 2024; Ahn et al., 2024; Hyung et al., 2025), a degraded or learned negative (Sadat et al., 2025; Han et al., 2026; Li et al., 2025b), or an appearance that cancels out of a motion (Huang et al., 2025); our negative is the reference itself with its transition replaced by a dissolve, so its scenes cancel and only the effect is amplified. Evaluation scores realism (Unterthiner et al., 2018), general quality (Huang et al., 2024b), or similarity to the reference (Radford et al., 2021; Zhao et al., 2024a; Yatim et al., 2024), and similarity rewards copying the input (Krojer et al., 2024). No existing dataset shows one two-sided transition effect on many endpoint pairs, and transition benchmarks score how clips connect without a reference effect (Yin et al., 2026).

Transition effect generation. Work on transitions connects two given endpoints, but the transition is chosen by the model or by text, never read from a reference clip. Inbetweening bridges two frames (Reda et al., 2022; Feng et al., 2024; Xing et al., 2024; Wang et al., 2025; Zhang et al., 2026), transition generation steers the bridge with text or structure (Chen et al., 2023; Yang et al., 2025; Kan et al., 2026), and video foundation models take a first and a last frame (Wan et al., 2025; HaCohen et al., 2026); several accept clips at both ends (Zhang et al., 2024; Kan et al., 2026; Jiang et al., 2025). Commercial tools pick from a preset library (Higgsfield AI, 2026; Shen et al., 2022), and

morphing (Beier & Neely, 1992) and transition detection (Zhang et al., 1993; Soucek & Lokoc, 2024) predate generative models. The task that does read an effect from a reference is visual effect transfer: it applies the effect of a reference video to a single start image and leaves the end open (Li et al., 2025a; Bian et al., 2026; Jones et al., 2026; Yang et al., 2026a; Ma et al., 2026). Most of these methods also give the model text about the effect (refVFX makes it optional): a prompt that names it (Li et al., 2025a; Bian et al., 2026; Ma et al., 2026), a description a multimodal LLM writes from the reference (Yang et al., 2026a), or a prompt optimized to match the reference (Zhao & Shou, 2026). Others generate effects from text alone (Mao et al., 2026; Liu et al., 2025; Yang et al., 2026b) or copy pixels of the reference into the target (Guo et al., 2024a). SEGUE receives no text about the effect, reads it from the reference alone, and must arrive at a given end clip.

3 SEGUE: Isolating and Amplifying the Reference Transition

SEGUE addresses TEG through two complementary mechanisms (Figure 3): training-side *reference isolation*, which combines neutral prompts with a one-way attention mask so the model reads the effect strictly from the visual reference (§3.2), and inference-side *null-reference guidance* to amplify that effect during inference (§3.3).

3.1 One-Shot Transition Effect Generation (TEG) Formulation

TEG transfers a transition effect from a visual reference clip V_r (connecting scenes A_r and B_r) onto target endpoints A and B in a one-shot manner. Given a text prompt c, the model generates a video:

$$V = f_\theta(V_r, A, B, c). \tag{1}$$

The output must satisfy three core criteria: (i) *transition fidelity*, faithfully transferring the reference transition to the generated transition; (ii) *endpoint fidelity*, opening seamlessly on A and closing on B, preserving their visual content and motion; and (iii) *reference disentanglement*, preventing the reference's original scenes (A_r, B_r) from leaking into the output. Each endpoint flexibly accommodates a video clip, a static image, or nothing ($B = \varnothing$), unifying two-sided TEG and single-sided visual effect transfer within a single architecture.

3.2 Training: Visual Reference Isolation

We structure training around *visual reference isolation*: the model reads the reference through a neutral prompt and a one-way attention mask, so the prompt says nothing about the effect and the target never rewrites the reference. This closes the text shortcut that previous approaches take (Figure 2) and leaves the reference as the only input that carries the effect. Below we describe the model (Figure 3): its token layout and objective, the attention mask, and the neutral prompt.

In-context reference conditioning. Let $V = [A, M, B]$ denote the target video, where the middle M holds the frames that carry the transition between the endpoints A and B. Let $\mathbf{z}_r$, $\mathbf{z}_A$, $\mathbf{z}_M$, and $\mathbf{z}_B$ be the VAE latents of the reference, the two endpoints, and the middle. Following in-context conditioning for diffusion transformers (Huang et al., 2024a), we place the reference, the endpoints, and the noised middle in one token sequence on a single positional grid:

$$x_\sigma = \big[\, \mathbf{z}_r,\ \mathbf{z}_A,\ \mathbf{z}_M^\sigma,\ \mathbf{z}_B \,\big], \qquad \mathbf{z}_M^\sigma = (1-\sigma)\mathbf{z}_M + \sigma\epsilon, \tag{2}$$

where $\sigma \in [0, 1]$ is the noise level, and the diffusion transformer reads x_σ as one sequence of tokens. Only the middle is noised: $\mathbf{z}_r$, $\mathbf{z}_A$, and $\mathbf{z}_B$ are given without noise and serve as context. This sequential layout handles every endpoint type without a dedicated conditioning branch, in the spirit of unified video conditioning (Jiang et al., 2025): a clip contributes all of its frames, an image contributes one frame, and an absent endpoint contributes no tokens. We build the training data with different endpoint settings, so one set of adapter weights covers two-sided transitions between video clips or static images as well as single-sided visual effect transfer from a start alone. We train LoRA adapters (Hu et al., 2022) on a frozen video diffusion backbone, with the flow-matching objective (Lipman et al., 2022) restricted to the middle:

$$\mathcal{L}(\theta) = \mathbb{E}_{\sigma,\,\epsilon}\Big[\, \big\| \mathbf{m} \odot \big(u_\theta(x_\sigma, c) - u\big) \big\|^2 \Big], \qquad \mathbf{m}_i = \mathbf{1}\{\, i \in M \,\}, \tag{3}$$

where u is the target velocity, c is the text prompt of Eq. (5), and $\mathbf{m}$ masks the loss outside M. With fixed endpoints, all training signal is focused on the transition.

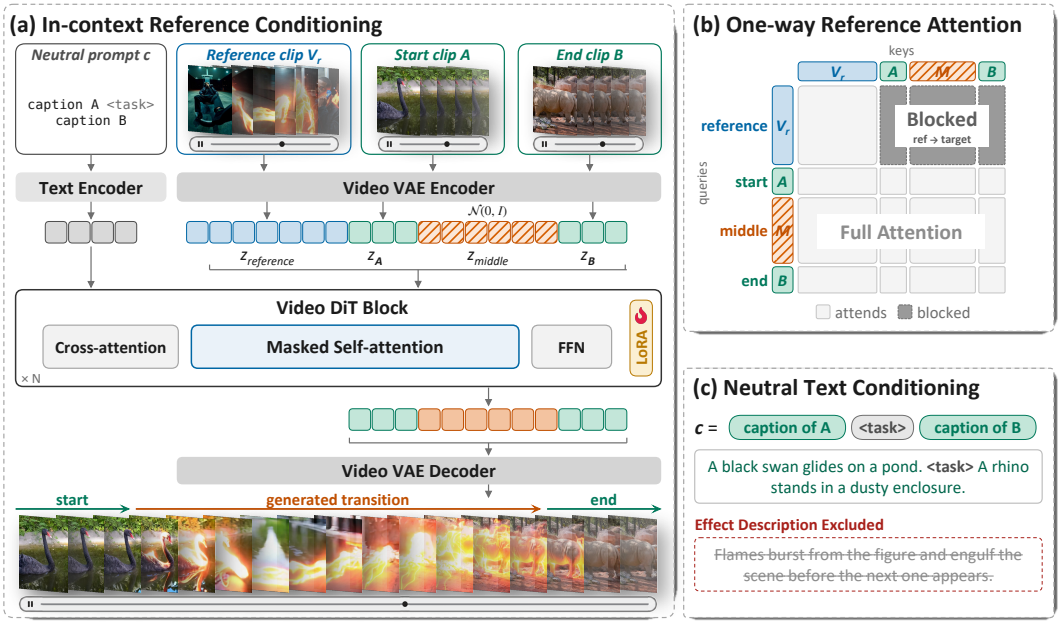

Figure 3: **SEGUE overview.** (a) In-context reference conditioning: the reference V_r, start A, and end B are encoded with the noised middle in one sequence, which the LoRA-adapted DiT denoises; the neutral prompt c enters through cross-attention. (b) One-way reference attention: masked self-attention blocks the reference from attending to the target and keeps full attention elsewhere. (c) Neutral text conditioning includes only captions of A and B without the effect description.

One-way attention mask. With full self-attention, the reference tokens would also read the target tokens, and layer by layer the representation of the reference would drift toward the target scenes. We therefore make the attention one-way. Let R index the reference tokens and G the remaining tokens, the endpoints and the middle. The mask is (Figure 3b):

$$\text{token } i \text{ attends to token } j \iff i \in G \text{ or } j \in R. \tag{4}$$

The target tokens G read the reference R, but the reference tokens never read G. The representation of the reference therefore depends on V_r alone, in every layer, and the target cannot alter it. This one-way rule is adapted from the reference attention of VFXMaster (Li et al., 2025a), which introduced it for effect transfer onto a single frame; their mask also covers text tokens and a reference prompt, neither of which is in our sequence (refer to Appendix B).

Neutral text conditioning. The text shortcut runs through the prompt: if the prompt named the effect, the model could read it from there and ignore the motion in the reference. We therefore use a neutral prompt that describes only the two scenes (Figure 3c):

$$c = \big[\text{caption of } A\big]\ \langle\text{task}\rangle\ \big[\text{caption of } B\big], \tag{5}$$

where $\langle\text{task}\rangle$ is a fixed separator token. No word about the effect enters the prompt during training. This leaves the reference as the only source of the transition effect, a clean visual signal that inference can amplify (§3.3).

3.3 Inference: Null-Reference Guidance

Reference isolation makes the reference the carrier of the transition effect. At inference, we want a guidance that amplifies the transition effect without reference scene leakage.

Negative reference. Classifier-free guidance (Ho & Salimans, 2022) amplifies a condition by steering away from the prediction without it, or away from a negative condition (Ban et al., 2024). With the reference as the condition, this reads $\hat{v} = v(\varnothing) + w\,\big(v(V_r) - v(\varnothing)\big)$. But a reference clip carries both, the transition effect and its scenes (A_r, B_r); and an empty reference removes both. The contrast therefore amplifies the scenes of the reference along with its effect, which is the reference scene leakage we want to avoid (Figure 13 in Appendix E). Instead, we want a negative reference that keeps the reference's scenes and removes only the effect.

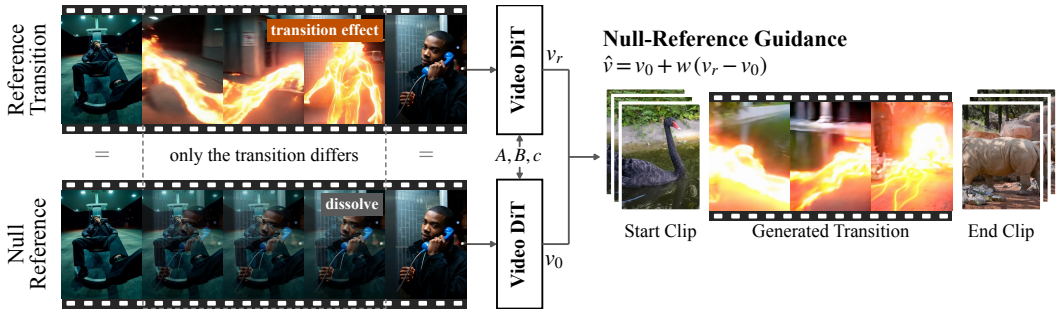

Figure 4: **Null-reference guidance.** The null reference V_r^0 replaces the transition in V_r with a dissolve between the same reference endpoints. The model processes both with identical endpoints (A, B) and prompt c, so their predictions, v_r with the reference and v_0 with the null reference, differ primarily in the transition effect; guidance steers away from the dissolve and toward the effect.

The null transition. We observe such a condition in the base model. Given two endpoints and a prompt that describes only the two scenes, LTX-2 (HaCohen et al., 2026) mostly produces a straight cut or a dissolve (Appendix G): it keeps both scenes and adds no motion and no material of its own. We take the dissolve, a linear blend of two frames, as the *null transition*,

$$T_0(a, b): \quad V_\tau = (1 - \tau)\, a + \tau\, b, \qquad \tau \in [0, 1], \tag{6}$$

where a is the last frame of the start and b the first frame of the end. The *null reference* is the reference with its own transition replaced by the null transition:

$$V_r^0 = \big[\, A_r,\ T_0(A_r, B_r),\ B_r \,\big]. \tag{7}$$

V_r and V_r^0 show the same scenes and differ only in the transition, which is the effect in V_r and a dissolve in V_r^0 (Figure 4).

Null-reference guidance (NRG). Like classifier-free guidance, at each sampling step we run the model twice, with the same endpoints (A, B), the same prompt c, and the same noise, and change only the reference. With $v_r = u_\theta\big(x_\sigma(V_r), c\big)$ and $v_0 = u_\theta\big(x_\sigma(V_r^0), c\big)$,

$$\hat{v} = v_0 + w\,(v_r - v_0), \qquad w \geq 1. \tag{8}$$

At $w = 1$ this is the model without guidance; for $w > 1$ the sample moves away from the dissolve and toward the effect of the reference. Everything except the effect is shared by the two passes, so the contrast is specific to the effect. The null reference is a valid video and not an empty input, so no condition dropout is needed in training and NRG applies to the trained model as it is.

3.4 SEGUE-BENCHMARK

Existing video datasets rarely show one transition effect on many scenes: transitions in available footage are one-off instances tied to specific clips. We therefore build the SEGUE-Benchmark, in which every transition effect is a class with many videos. A training sample pairs two videos of the same class on different endpoints: one is the reference V_r, and the other is the target V, whose endpoints (A, B) are given as conditions. The data is *cross-paired*: each transition effect appears on many endpoint pairs, and each endpoint pair under many transition effects, so effects and endpoints are decorrelated and neither can be inferred from the other.

Table 1: **Data sources.** A training pair is a target video with up to three references of the same effect. External effects include the 34 EffectData effects, used only for evaluation.

Source	Effect types	Distinct endpoints	Videos	Training pairs	Effects per endpoint	Endpoints per effect
Creative transitions	26	350	1,364	4,060	3.6	49.0
Procedural transitions	1,590	1,103	15,436	46,308	14.0	9.7
External effects	76	2,051	2,034	6,000	1.0	27.2
Total	1,692	3,504	18,834	56,368	5.4	11.1

The training set combines three sources (Table 1), chosen to cover many kinds of transitions and to keep the pairing dense, with many effects per endpoint and many endpoints per effect. (i) *Creative transitions*: 26 preset transitions from Higgsfield (Higgsfield AI, 2026), each learned by a small LoRA and rendered on a shared set of endpoints, with every video inspected by hand. (ii) *Procedural*

Table 2: **Evaluation metrics.** V: the generated video; V_r: the reference clip; $\mathcal{R}$: real videos of the reference's transition class. Full definitions in Appendix C.

Metric	What it checks	Score
Transition fidelity ↑	Does V show the transition effect of V_r?	$100 \cdot \mathrm{sim}(V, \mathcal{R}) \,/\, \mathrm{sim}(\mathcal{R}, \mathcal{R})$; sim: change over time (DINOv2), motion (CoTracker3), appearance; real videos score 100
Motion fidelity ↑	Does V move like V_r?	$100 \cdot \mathrm{corr}\big(\mathrm{Tracks}(V), \mathrm{Tracks}(V_r)\big)$, CoTracker3 point tracks
Action KL ↓	Does V show the same action as V_r?	$\mathrm{KL}\big(\mathrm{Swin}(V_r) \,\|\, \mathrm{Swin}(V)\big)$, Video Swin action probabilities
Reference similarity ↑	Does V look like V_r?	$100 \cdot \cos\big(\mathrm{VideoPrism}(V), \mathrm{VideoPrism}(V_r)\big)$
Seam-free rate ↑	Does V leave the given clips without a visible cut?	% of outputs whose LPIPS change at the hand-off is within their usual frame-to-frame change

transitions: 1,590 parametric transitions, such as geometric wipes, volumetric zooms, fluid warps, and alpha-blended shaders with randomized direction, speed, and boundary softness, rendered on shared endpoints. (iii) *External visual effects* from prior visual effect transfer work, refVFX (Jones et al., 2026) and EffectData (Yang et al., 2026a), which add single-sided visual effect transfer and other endpoint configurations. In total, the benchmark has 1,692 effect types and 18,834 videos. The 34 EffectData effects and 17 classes of our own sources are held out of training and form the evaluation dataset (Section 4.1).

4 Experiments

4.1 Experimental Setup

Implementation details. SEGUE is a rank-128 LoRA adapter (Hu et al., 2022) on the frozen LTX-2 19B model (HaCohen et al., 2026), trained on 8 H100 GPUs on the 56K training pairs of Table 1. Outputs are 480×640 at 24 fps and 121 frames long. We sample with 30 steps and null-reference guidance at $w = 6$ in every table; Appendix E sweeps w and Appendix B gives details.

Evaluation dataset and metrics. The evaluation dataset of SEGUE-Benchmark is fully held out: neither its transitions nor its endpoints appear in training. Each evaluation sample consists of a reference clip that shows a transition effect and the target endpoints to which the effect must be transferred: a start and an end for two-sided TEG (38 samples), or the start alone for single-sided visual effect transfer (183 samples). The 221 samples cover 51 transition classes (the 34 EffectData classes and 17 classes held out from our own sources) and 150 distinct endpoints, and each sample is generated with two seeds. Appendix F lists the classes and the samples. We report five metrics, defined in Table 2. *Transition fidelity* compares the output with real videos of the reference's transition class in change over time, motion, and appearance, scaled so that real videos score 100. *Motion fidelity*, *action KL*, and *reference similarity* compare the output with the reference clip in tracked motion, action class, and appearance, and the *seam-free rate* checks the hand-off to the given clips for a visible cut. Every system is run with two prompts: the neutral prompt (N), which describes only the endpoints, as in Eq. (5), and the effect prompt (E), which also describes the reference transition. For every metric we report both scores and $|\Delta|$, the absolute difference between them. A small $|\Delta|$ means that the model reads the transition mainly from the reference and depends little on the text prompt.

Baselines. No previous approach is designed for a reference clip together with two endpoints. For two-sided TEG, we therefore adapt existing models as baselines: the base LTX-2 (HaCohen et al., 2026), VACE (Jiang et al., 2025), and Wan2.1 FLF2V (Wan et al., 2025) take both endpoints but no reference, so the effect prompt is their only source of the effect; refVFX (Jones et al., 2026) takes the reference and the start endpoint, and we adapt it by giving it the end endpoint as well. For single-sided visual effect transfer, we compare with the previous visual effect transfer methods Video-As-Prompt (Bian et al., 2026), VFXMaster (Li et al., 2025a), and refVFX, which take the reference and the start endpoint. Each runs in its native setting, with its own prompt format, resolution, and frame count.

4.2 Analysis

SEGUE transitions are more faithful. To test whether a model can transfer a transition effect it has never seen, we compare SEGUE with previous approaches on held-out effects, in both tasks with the neutral and the effect prompt (Table 3). SEGUE produces the most faithful transitions in both tasks: with the neutral prompt, it scores close to real videos of the same effect, while previous approaches reach about half of that. Even with the effect prompt, they stay below SEGUE, which

Table 3: **Comparison with previous approaches and adapted baselines** on the evaluation dataset. N and E: the neutral prompt, which describes only the endpoints, and the effect prompt, which also describes the reference transition; $|\Delta|$: the absolute gap between them. SEGUE w/o NRG: no guidance ($w = 1$), not ranked; small numbers under its values: its difference from SEGUE (green: better, red: worse).

	Transition fidelity ↑			Motion fid. (×100) ↑			Action KL ↓			Ref sim. (×100) ↑			Seam-free % ↑		
	N	E	$\|\Delta\|$	N	E	$\|\Delta\|$	N	E	$\|\Delta\|$	N	E	$\|\Delta\|$	N	E	$\|\Delta\|$
Transition effect generation: both endpoints given															
Base LTX-2	49.8	84.7	34.9	8.8	8.1	0.7	5.98	5.48	0.50	94.8	95.3	0.5	80.3	93.4	13.1
VACE	51.9	77.0	25.1	11.9	9.9	2.0	5.71	5.04	0.67	95.1	95.2	**0.1**	92.1	93.4	**1.3**
refVFX	49.8	70.5	20.7	**14.1**	11.4	2.7	5.68	5.38	0.30	95.2	95.4	0.2	88.2	97.4	9.2
Wan2.1 FLF2V	56.6	84.3	27.7	9.4	7.4	2.0	5.59	4.98	0.61	95.0	95.2	0.2	94.7	**100.0**	5.3
SEGUE	**97.0**	**97.9**	**0.9**	12.8	**12.5**	**0.3**	**4.62**	**4.52**	**0.10**	**95.9**	**96.0**	**0.1**	**97.4**	98.7	**1.3**
SEGUE w/o NRG	92.9 −4.1	96.2 −1.7	3.3 +2.4	11.7 −1.1	12.0 −0.5	0.3	5.03 +0.41	5.14 +0.62	0.11 +0.01	95.8 −0.1	95.9 −0.1	0.1	100.0 +2.6	100.0 +1.3	0.0 −1.3
Visual effect transfer: start endpoint given															
Video-As-Prompt	40.9	70.3	29.4	12.9	13.7	0.8	5.69	4.41	1.28	94.7	95.8	1.1	92.1	93.2	1.1
VFXMaster	49.9	85.5	35.6	17.5	19.2	1.7	5.48	4.50	0.98	94.9	96.2	1.3	**96.7**	**99.7**	3.0
refVFX	42.1	66.3	24.2	18.0	16.1	1.9	5.60	4.75	0.85	94.9	95.7	0.8	66.7	94.5	27.8
SEGUE	**92.6**	**94.0**	**1.4**	**22.6**	**22.7**	**0.1**	**4.15**	**4.05**	**0.10**	**96.5**	**96.6**	**0.1**	93.4	92.9	**0.5**
SEGUE w/o NRG	86.4 −6.2	91.0 −3.0	4.6 +3.2	20.3 −2.3	20.4 −2.3	0.1	4.52 +0.37	4.29 +0.24	0.23 +0.13	96.3 −0.2	96.5 −0.1	0.2 +0.1	96.4 +3.0	96.2 +3.3	0.2 −0.3

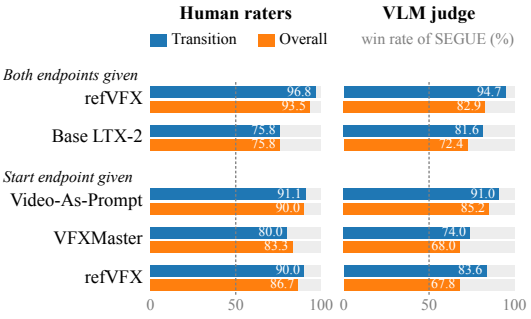

Figure 5: **Human study and VLM judge.** Win rate (%) against each system for transition and overall quality. Protocol in Appendix D.

Table 4: **Flagged clips** (%) by the raters' majority and by the VLM judge, for a missing effect, a missed endpoint (End., among clips with an effect), and leaked reference content.

	Raters (%) ↓			VLM judge (%) ↓		
	Effect	End.	Leak	Effect	End.	Leak
Both endpoints given						
refVFX	40.3	41.9	19.4	5.3	8.3	0.0
Base LTX-2	40.3	40.3	6.5	10.5	2.9	–
SEGUE	0.0	1.6	4.8	0.0	5.3	5.3
Start endpoint given						
Video-As-Prompt	32.2	35.6	23.3	4.4	0.0	1.1
VFXMaster	24.4	30.0	14.4	0.0	0.0	0.0
refVFX	26.7	32.2	16.7	1.1	0.6	0.0
SEGUE	0.0	1.1	1.1	2.2	0.0	0.0

shows that a text description cannot convey the motion and timing that the reference clip shows. The other reference metrics agree, and the seam-free rate shows that the hand-off to the given clips stays clean. The qualitative comparison makes the difference visible (Figure 6; Figure 8 in Appendix E for the start alone): SEGUE reproduces the dynamics, timing, and appearance of the reference between the endpoints without leaking its scenes, whereas previous approaches render a generic frame-filling effect even with the reference or its description.

SEGUE uses the visual signal of the reference transition. Our hypothesis is that previous approaches take the effect from the text prompt rather than from the reference. To measure this, we run every system with the effect prompt (E) and with the neutral prompt (N) and report the gap $|\Delta|$ (Table 3). Previous approaches lose most of the effect once the prompt stops describing it, so their visual reference contributes little. SEGUE barely changes on any metric: it reads the transition from the reference clip, and the prompt adds almost nothing to what the reference already carries.

Human study and VLM judge. Automatic metrics may miss how people perceive a transition, so we also ask human raters and a VLM judge to choose between SEGUE and one other model, on the transition and overall, and to flag clips with a missing effect, missed endpoint, or leaked reference content (Figure 5 and Table 4; protocol in Appendix D). With both endpoints given, the opponents are refVFX and the base LTX-2 with the effect prompt; with the start alone, Video-As-Prompt, VFXMaster, and refVFX. Raters and the judge prefer SEGUE over every other model, significantly on both transition fidelity and overall quality. The flags follow the same order; the judge is more lenient than the raters on whether an effect is present.

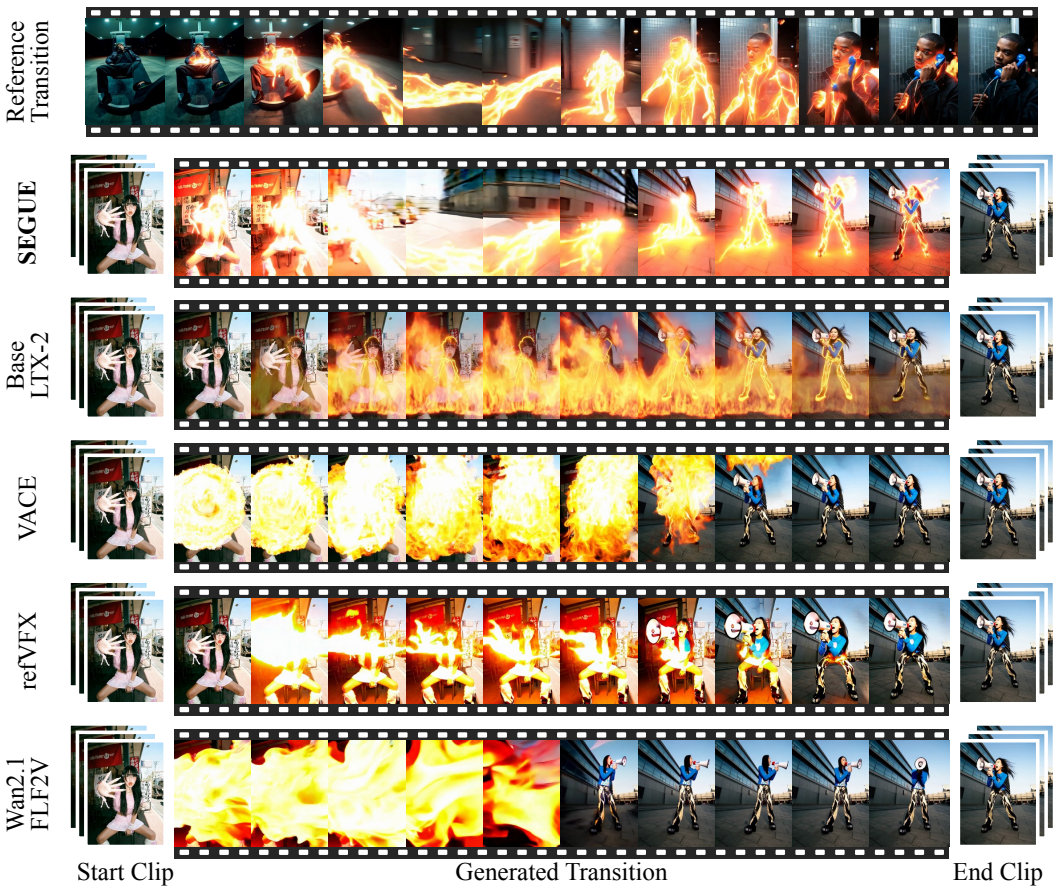

Figure 6: **Qualitative comparison**, both endpoints given. Top: the reference; below: one system per row, frames in time order. Videos and additional results are available in the supplemental.

Ablations. To isolate the role of null-reference guidance, we run SEGUE without it ($w = 1$), the SEGUE w/o NRG rows of Table 3. Without guidance, the transitions are less faithful to the reference in both tasks and depend more on the effect prompt, while the hand-off becomes slightly smoother: guidance strengthens the effect that the reference carries, at a small cost in seamlessness. Appendix E repeats the ablation on references whose transition classes were seen in training and sweeps the guidance weight; we use $w = 6$ throughout.

5 Conclusion

We introduced One-Shot Transition Effect Generation (TEG) and SEGUE, a model that transfers a transition effect from a single reference clip to new endpoints, whether video clips, static images, or absent. SEGUE combines visual reference isolation during training with null-reference guidance at inference, so the transition is read and amplified from the visual demonstration rather than from the effect prompt. We also introduced the SEGUE-Benchmark, a cross-paired dataset with a held-out evaluation dataset and the transition fidelity metric. Our results show that a visual reference alone is sufficient to specify and reuse a transition effect across endpoints.

Limitations. Two failure modes appear on a few samples, shown in Figure 14 of Appendix E. As the guidance weight grows, on some samples we observed scene leakage from the reference, and for one specific endpoint that shows intense motion, the endpoints were not preserved. As any instance of classifier-free guidance, our null-reference guidance doubles the model evaluations per sampling step, and all results are on one base model, LTX-2 (HaCohen et al., 2026).

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic training data: the creative and procedural transition videos of the SEGUE-Benchmark are rendered by video generation models (Section 3.4, Appendix A), and the scene descriptions in the training prompts are written by a vision-language model from the endpoint frames (Appendix A). A vision-language model also serves as the automatic judge in our evaluation (Section 4.2, Appendix D); its outputs are reported as measurements and are compared with human raters. We did not use generative AI tools to develop the ideas, to propose or refine hypotheses, to design the method or the experiments, to implement the method, or to interpret the results. Theoretical claims and proofs are not applicable to this work. Additionally, we used AI tools to search for and identify related work, after which the authors read every cited paper, to help write auxiliary code such as data processing and plotting scripts, and to edit the text of this paper for readability. We have reviewed all AI-assisted work and take full responsibility for the content of this paper.

ETHICS STATEMENT

We have read and adhere to the ICLR Code of Ethics. The human study (Appendix D) was fully anonymous: raters only compared generated videos, and no personal information was collected. The endpoint clips come from public video datasets and from generated videos, and some of them show people; we do not annotate or infer any attribute of any person. Third-party effects and datasets are used under their licenses, and EffectData is used for evaluation only. The model joins two clips that the user already has and leaves both unchanged, which limits its use for impersonation; like any video generator, however, it could make edited footage look more convincing.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B describe the model, the training setup, and the sampler. Sections 3.4 and 4.1 and Appendices A and F describe the training data and the evaluation dataset; Appendix C defines every metric, and Appendix D gives the protocol of the human study. Previous methods and adapted baselines are run with their released code. We will release the code, the LoRA weights, the evaluation dataset, and the implementation of the metrics.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *european conference on computer vision*, pp. 190–206. Springer, 2024.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in neural information processing systems*, 35:25005–25017, 2022.

Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. In *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 35–42, 1992. doi: 10.1145/133994.134003.

Yuxuan Bian, Xin Chen, Zenan Li, Tiancheng Zhi, Shen Sang, Linjie Luo, and Qiang Xu. Video-as-prompt: Unified semantic control for video generation. In *International Conference on Learning Representations*, volume 2026, pp. 114558–114583, 2026.

Lan Chen, Qi Mao, Yuchao Gu, and Mike Zheng Shou. Edit transfer: Learning image editing via vision in-context relations. *arXiv preprint arXiv:2503.13327*, 2025.

Tianxiao Chen, Hanmo Chen, Huajin Chen, Bo Li, Qi Ye, and Peng-Tao Jiang. Viper: Visual in-context physics reasoning for physically plausible video generation. *arXiv preprint arXiv:2607.23472*, 2026.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023.

Zhengcong Fei, Di Qiu, Debang Li, Changqian Yu, and Mingyuan Fan. Video diffusion transformers are in-context learners. *arXiv preprint arXiv:2412.10783*, 2024.

Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *European Conference on Computer Vision*, pp. 378–395. Springer, 2024.

GL Transitions contributors. GL transitions: a collection of GLSL transitions. https://gl-transitions.com, 2026. Accessed 2026-09.

Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Transactions on Graphics (TOG)*, 43(4): 1–15, 2024.

Jiaqi Guo, Lianli Gao, Junchen Zhu, Jiaxin Zhang, Siyang Li, and Jingkuan Song. Magicvfx: Visual effects synthesis in just minutes. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8238–8246, 2024a.

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*, pp. 330–348. Springer, 2024b.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Shilong Han, Yuming Zhang, and Hongxia Wang. Guiding diffusion models with semantically degraded conditions. *arXiv preprint arXiv:2603.10780*, 2026.

Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 557–570. 2023.

Higgsfield AI. Higgsfield: Transition app. https://higgsfield.ai, 2026. Accessed 2026-09.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. arXiv:2106.09685.

Chi-Pin Huang, Yen-Siang Wu, Hung-Kai Chung, Kai-Po Chang, Fu-En Yang, and Yu-Chiang Frank Wang. Videomage: Multi-subject and motion customization of text-to-video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17603–17612. IEEE, 2025.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024a.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024b.

Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11006–11015. IEEE, 2025.

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

Maxwell Jones, Rameen Abdal, Or Patashnik, Ruslan Salakhutdinov, Sergey Tulyakov, Jun-Yan Zhu, and Kuan-Chieh Jackson Wang. Tuning-free visual effect transfer across videos. In *European Conference on Computer Vision*, pp. 654–674. Springer, 2026.

Mia Kan, Yilin Liu, and Niloy J Mitra. Sage: Structure-aware generative video transitions between diverse clips. In *Computer Graphics Forum*, pp. e70420. Wiley Online Library, 2026.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labeling real videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Christopher Pal, and Siva Reddy. Learning action and reasoning-centric image editing from videos and simulation. *Advances in Neural Information Processing Systems*, 37:38035–38078, 2024.

Baolu Li, Yiming Zhang, Qinghe Wang, Liqian Ma, Xiaoyu Shi, Xintao Wang, Pengfei Wan, Zhenfei Yin, Yunzhi Zhuge, Huchuan Lu, et al. Vfxmaster: Unlocking dynamic visual effect generation via in-context learning. *arXiv preprint arXiv:2510.25772*, 2025a.

Xiaomin Li, Yixuan Liu, Takashi Isobe, Xu Jia, Qinpeng Cui, Dong Zhou, Dong Li, You He, Huchuan Lu, Zhongdao Wang, et al. Reneg: Learning negative embedding with reward guidance. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23636–23645. IEEE, 2025b.

Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. Visualcloze: A universal image generation framework via visual in-context learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18969–18979. IEEE, 2025c.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xinyu Liu, Ailing Zeng, Wei Xue, Harry Yang, Wenhan Luo, Qifeng Liu, and Yike Guo. Vfx creator: Animated visual effect generation with controllable diffusion transformer. *arXiv preprint arXiv:2502.05979*, 2025.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3192–3201. IEEE, 2022.

Yue Ma, Xu Ye, Qinghe Wang, Yucheng Wang, Hongyu Liu, Yinhan Zhang, Xinyu Wang, Yuanpeng Chen, Shanhui Mo, Paul Liang, et al. Easyvfx: frequency-driven decoupling for resource-efficient vfx generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2026.

Fangyuan Mao, Aiming Hao, Jintao Chen, Dongxia Liu, Xiaokun Feng, Jiashu Zhu, Meiqi Wu, Chubin Chen, Jiahong Wu, and Xiangxiang Chu. Omni-effects: Unified and spatially-controllable visual effects generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7927–7935, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision*, pp. 250–266. Springer, 2022.

Thomas Ressler-Antal, Frank Fundel, Malek Ben Alaya, Stefan Andreas Baumann, Felix Krause, Ming Gui, and Björn Ommer. Dismo: Disentangled motion representations for open-world motion transfer. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 158494–158534. Curran Associates, Inc., 2025. doi: 10.52202/085713-5296. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/e88870ec82f2469b0ddf32c817920c68-Paper-Conference.pdf.

Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 76833–76858, 2025.

Yaojie Shen, Libo Zhang, Kai Xu, and Xiaojie Jin. Autotransition: Learning to recommend video transition effects. In *European Conference on Computer Vision*, pp. 285–300. Springer, 2022.

Tomás Soucek and Jakub Lokoc. 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, 2024.

Yihong Sun, Hao Zhou, Liangzhe Yuan, Jennifer J Sun, Yandong Li, Xuhui Jia, Hartwig Adam, Bharath Hariharan, Long Zhao, and Ting Liu. Video creation by demonstration. *arXiv preprint arXiv:2412.09551*, 2024.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steve Seitz. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. In *International Conference on Learning Representations*, volume 2025, pp. 30594–30609, 2025.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6830–6839. IEEE, 2023a.

Yuji Wang, Teng Hu, Yuheng Chen, Ran Yi, Han Feng, Weijian Cao, Chengjie Wang, Lizhuang Ma, and Jiangning Zhang. Vicedit: Learning to edit videos from visual in-context examples. *arXiv preprint arXiv:2608.16745*, 2026.

Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems*, 36:8542–8562, 2023b.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. Humanvid: Demystifying training data for camera-controllable human image animation. *Advances in Neural Information Processing Systems*, 37:20111–20131, 2024a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dream video: Composing your dream videos with customized subject and motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6537–6549. IEEE, 2024.

Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *ACM Transactions on Graphics (TOG)*, 43(6):1–11, 2024.

Xuancheng Xu, Gengyun Jia, and Bing-Kun Bao. Disco-lora: Disentangled composition of content, style, and motion for multi-concept video customization. *arXiv preprint arXiv:2606.26668*, 2026.

Shiyuan Yang, Ruihuang Li, Jiale Tao, Shuai Shao, Qinglin Lu, and Jing Liao. Effectmaker: Unifying reasoning and generation for customized visual effect creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16108–16117, 2026a.

Xiaomeng Yang, Yanyu Li, Gordon Guocheng Qian, Ivan Skorokhodov, Viacheslav Ivanov, Avalon Vinella, Xuan Zhang, Yanzhi Wang, Sergey Tulyakov, and Anil Kag. Prompt2effect: Training-free image-to-video model specialization via lora generation. In *European Conference on Computer Vision*, pp. 482–499. Springer, 2026b.

Zuhao Yang, Jiahui Zhang, Yingchen Yu, Shijian Lu, and Song Bai. Versatile transition generation with image-to-video diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16981–16990. IEEE, 2025.

Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8466–8476. IEEE, 2024.

Zhiyu Yin, Zhipeng Liu, Kehai Chen, Lemao Liu, Jin Liu, Hong-Dong Li, Yang Xiang, and Min Zhang. Vc-bench: Pioneering the video connecting benchmark with a dataset and evaluation metrics. *arXiv preprint arXiv:2601.19236*, 2026.

Bowen Zhang, Xiaofei Xie, Haotian Lu, Na Ma, Tianlin Li, and Qing Guo. Mavin: Multi-action video generation with diffusion models via transition video infilling. *arXiv preprint arXiv:2405.18003*, 2024.

HongJiang Zhang, Atreyi Kankanhalli, and Stephen W Smoliar. Automatic partitioning of full-motion video. *Multimedia systems*, 1(1):10–28, 1993.

Lin Zhang, Sicheng Mo, Zefan Cai, Jinhong Lin, Zihao Lin, Jiuxiang Gu, Krishna Kumar Singh, Yuheng Li, and Yin Li. Unitemp: Unlocking video generation in any temporal order via bidirectional distillation. *arXiv preprint arXiv:2606.18702*, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217*, 2024a.

Rui Zhao and Mike Zheng Shou. P-flow: Prompting visual effects generation. *arXiv preprint arXiv:2603.22091*, 2026.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 273–290. Springer, 2024b.

SUPPLEMENTARY MATERIAL

CONTENTS

A Dataset construction **17**

B Training details **18**

C Evaluation metrics **19**

D Human study and VLM judge protocol **20**

E Additional results **25**

F Evaluation dataset **33**

G Distance from linear interpolation **34**

A DATASET CONSTRUCTION

Creative transitions. We collect videos of 43 preset transitions from Higgsfield (Higgsfield AI, 2026), 26 for training and 17 held out, and standardise every clip to 121 frames at 24 fps and 480×640 pixels. To show one transition on many scenes, we train a small LoRA per training transition on its collected clips and render the transition onto 350 shared endpoints: the 139 Higgsfield scenes and 211 clips from DAVIS (Pont-Tuset et al., 2017) and HumanVid (Wang et al., 2024a). Every rendered clip is inspected by hand, and clips with a boundary artifact, a frozen or black frame, or a changed start scene are removed; 192 of 1,417 rendered clips were rejected. The row of Table 1 counts the 139 collected clips and the 1,225 kept renders.

Procedural transitions. Each procedural transition is a GLSL shader from the gl-transitions bank (GL Transitions contributors, 2026) with sampled parameters: the shader uniforms, an easing curve, the onset and release times, a flip, and a swap of the two inputs. The timing is part of the transition, so two samplings of one shader are two transitions. We render each transition between 10 pairs of real clips at 121 frames, 24 fps, and 480×640; the first and last 9 frames are the source clips byte for byte, so the endpoints are exact by construction. The endpoint clips come from DAVIS, HumanVid, and VC-Bench (Yin et al., 2026). Shader families are held out whole: 10 families are never rendered, each with a same-genre cousin in training.

External effects. From refVFX (Jones et al., 2026) we take all 42 effect LoRAs at their native 33 frames, 16 fps, and 832×448, one effect per clip and 47 to 50 clips per effect; 5 further refVFX effects are held out. From EffectData (Yang et al., 2026a) we select 34 effects as 17 couples that share a start subject, keep one reference video per effect and 51 start subjects, and never train on them.

Training pairs. Within each effect the clips are sorted, and every clip is the target of up to three pairs whose references are the next three clips of the same effect in cyclic order. The reference and the target never share an endpoint, and effects with fewer than two clips are dropped. This gives the 56,368 pairs of Table 1 from 18,800 videos.

Prompts. Every training prompt is rendered from per-clip scene descriptions by one template: the start scene, the trigger token, and, for a two-sided clip, the end scene, as in Eq. (5). The descriptions are written from the first nine frames of the clip, and for a two-sided clip also from the last nine, so no prompt describes the transition. A leak check confirms that a judge given only the prompt cannot identify the effect above chance.

Held-out sets. The held-out class tier of the evaluation dataset uses the 17 held-out Higgsfield transitions and the 34 EffectData effects. Its 150 endpoints and 68 reference clips are disjoint from the training set, including the DAVIS clips that the creative renders use.

B TRAINING DETAILS

Adapter. SEGUE is a rank-128 LoRA on the frozen LTX-2 19B model. The adapter sits on the query, key, value, and output projections of self- and cross-attention and on both feed-forward projections of every block, 654M parameters, 3.4% of the backbone. The backbone, its VAE, and its text encoder stay frozen.

Sequence layout. The token sequence of Eq. (2) holds the reference in full, followed by the target. A given start clip contributes its first 9 frames and a given end clip its last 8, a static image contributes one frame, and an absent endpoint contributes no tokens. The middle fills the rest of a 121-frame target, so a transition between two clips has 104 generated frames. All videos are 480×640 at 24 fps; the EffectData samples use 81-frame targets, the length of their source clips. The reference and the endpoints enter clean, only the middle is noised, and the loss covers only the middle (Eq. (3)).

Optimization. We train with AdamW at a learning rate of 10^{-4}, decayed linearly to 10^{-5} over 10,000 steps without warmup, in bf16 with an effective batch of 8 (8 H100 GPUs, one sample each), about 3 seconds per step. The endpoint setting is chosen per sample from the pairs of Table 1, so one adapter serves clips, images, and absent endpoints.

Sampling. Sampling uses 30 steps, the backbone's default text guidance of 4 and spatio-temporal guidance of 1, and null-reference guidance at $w = 6$ (Section 3.3). Null-reference guidance runs the model twice per step, once with the reference and once with the null reference, so a clip costs twice the time of the base model.

One-way reference attention. The rule of Eq. (4) is that of the reference attention of VFXMaster (Li et al., 2025a), applied to the video tokens: the reference tokens attend only to themselves, and the endpoint and middle tokens attend to everything. The mask of VFXMaster also covers the text tokens of its sequence and a reference prompt; LTX-2 reads the text through cross-attention and SEGUE uses no reference prompt, so no text token is masked. We implement the rule without a mask tensor, as two unmasked attention calls per layer, one for the reference queries over the reference keys and one for the target queries over all keys, and verified that the result matches the dense-mask attention. Both calls run on the fused scaled-dot-product attention kernels of PyTorch 2.9.1. For a forward and backward pass over 9,600 tokens in bf16 on an L40S GPU, attention takes 27.0 ms bidirectional, 82.5 ms with a dense mask, and 21.8 ms with the two calls, so the one-way rule is no slower than full attention.

C EVALUATION METRICS

This appendix defines the five metrics of the result tables, summarized in Table 2. Every metric compares the generated video V with the reference clip V_r, or with real videos of the transition effect that V_r shows. All are computed on the whole output at its native resolution and frame rate. Four of them sample the video at fixed fractions of its duration, so clips of different length and frame rate are compared on equal footing; the seam-free rate looks at consecutive frames instead.

Transition fidelity. The score asks whether V shows the same transition effect as the real videos of the reference's class, in five steps.

1. **Features.** Every frame is described by its DINOv2 class token (Oquab et al., 2023). The given windows at the two ends of the video, 9/121 and 8/121 of its length, give two endpoint anchors, and the frames between them are resampled to 32 steps of normalized time.
2. **Residual.** Each frame feature is projected off the plane spanned by the two anchors. What remains, the residual r_t, is the part of the frame that the endpoints do not explain: the effect.
3. **Three distances** between V and a real video U. *Change over time*: the residual steps $r_{t+1} - r_t$ are unit directions, each with a mass equal to its length and a time stamp equal to how far the clip has progressed from the start scene to the end scene; the two sets of steps are matched by optimal transport with cost $(1 - \cos) + 0.25\,|\Delta\text{progress}|$, so a mistimed or differently shaped change costs more. *Appearance*: every unit residual frame is matched to its most similar frame in the other video, in both directions and without regard to order; the distance is one minus the mean best cosine. *Motion*: CoTracker3 (Karaev et al., 2025) tracks a 20×20 grid of points; the tracks are smoothed and resampled to 32 steps, and each step is summarized by 18 numbers (fraction of moving points, mean speed, an 8-bin direction histogram weighted by speed, the divergence, curl, and two shear terms of an affine fit to the motion field, the mean translation, and the spread of the moving points), z-scored on the corpus; the distance is one minus the cosine of the two concatenated descriptors, so the order of the steps matters.
4. **Normalization.** Each distance becomes a rank: one minus its empirical CDF over all pairs among the 222 real transition clips of our sources. The three ranks are averaged with equal weights, and the average is ranked once more against the same average over the corpus pairs. This puts the three distances on one scale and makes the score comparable across effects.
5. **Score.** The similarity of V is averaged over the real videos of the reference's transition class, without the reference itself, and divided by the mean similarity of those real videos to one another. The result is multiplied by 100 and capped at 100, so a real video of the effect scores 100 on average.

Motion fidelity. As in Yatim et al. (2024): the CoTracker3 tracks of V and of V_r are reduced to unit velocity directions at 64 steps; every moving track is matched to the track of the other video with the highest direction correlation, and the mean over both directions is the score. A clip in which no point moves has no score and is left out of the mean.

Action KL. Thirty-two uniformly sampled frames of the whole video pass through Video Swin-B (Liu et al., 2022) trained on Kinetics-400. The score is the KL divergence from the class distribution of V_r to that of V: it is lowest when the output triggers the same action classes as the reference. The 400 classes serve as a signature of the motion, not as names of the effects.

Reference similarity. Cosine similarity between the VideoPrism (Zhao et al., 2024a) embeddings of V and of V_r, computed with the scorer released with refVFX, so the column is comparable with that paper.

Seam-free rate. LPIPS between consecutive frames, d_t, is computed over the whole output. At the hand-off from the given start frames to the generated frames, and, when both endpoints are given, at the hand-off back to the given end frames, the step d_t is compared with the other steps of the same video by a robust z-score (median and median absolute deviation). An output is seam-free when the larger of its hand-off z-scores is at most 3. The rate is the share of seam-free outputs.

Reporting. Each number in a result table is the mean over the videos of its block: 38 two-endpoint samples with two seeds each (76 videos) and 183 start-only samples with two seeds each (366 videos) on the held-out class tier; our own variants are also reported on the other tiers in Appendix E. Motion fidelity and reference similarity are printed times 100. $|\Delta|$ is the absolute difference between the two printed means, E and N. Bold marks the best value of each column within a block, separately for N and for E, and the smallest $|\Delta|$.

D HUMAN STUDY AND VLM JUDGE PROTOCOL

Samples and pairs. A pair shows the reference clip, the given start frame (and the given end frame for transition effect generation), and two outputs on the same sample: SEGUE and one other system, each at its native frame count and resolution. Samples come from the held-out class tier of the evaluation dataset (38 two-sided and 183 single-sided samples). The VLM judge rates every sample of the tier: the 38 two-sided samples against refVFX and Base LTX-2 (76 pairs), and the 183 single-sided samples against Video-As-Prompt, VFXMaster and refVFX (549 pairs); seed 42 throughout. The human study rates 197 of these pairs, on 76 samples chosen at random from the tier: 31 two-sided samples (62 pairs) and 45 single-sided samples, one per effect class (135 pairs).

Human raters. Raters were recruited on Prolific and the study is anonymous: the page asks for no name or identifier, and only the answers are recorded. A rater first reads a short instruction screen, then walks through four worked examples with their answers, and then rates a set of 15 pairs, about 8 minutes. Each pair is shown as in Figure 7: on the left the given start image (and the given end image for transition effect generation) and the reference, labelled "intended transition"; on the right the two outputs, labelled A and B, with the side of SEGUE drawn at random per pair. Every clip plays at its own native frame rate and length, holds its last frame when it ends, and all clips of a pair restart together; the instructions state that a shorter or less smooth video is never a reason to prefer one. Five questions follow in one row. Two are forced choices between A and B with no tie: (1) which video reproduces the intended transition on the given image(s) better, and (2) which video is the better effect-transfer result overall, considering the transfer of the intended transition, visual quality and aesthetics. Three are statements answered True or False for A and for B separately: (3) a real transition happens, not just a cut, fade or frozen image; (4) it starts on the start image (and ends on the end image); (5) it leaks content from the intended transition, where a leak is people, objects or background of the reference's own scene that are not part of the effect. A rater may go back to an earlier pair of the set and change an answer; the last answer counts. Pairs are served least-rated first, so every pair reaches three raters before any pair gets a fourth. Nothing on the page names a system, effect or prompt, and media files carry hashed names.

Scoring. A rater's own answer to statement (3) gates that rater's two choices on the pair: if only one of the two clips shows a transition, that clip wins both choices; if neither does, the pair is a tie for that rater; otherwise the rater's choices stand. Per pair, the gated votes are combined by majority (an even split is a tie), and the win rate of SEGUE against a system is the mean over that system's pairs with a tie counted as half. Significance is a two-sided sign test on the decided pairs; the interval is a Wilson 95% interval over pairs. Flag rates are per clip: a clip is flagged when the majority of its raters answer False to statement (3) or (4), or True to statement (5); a clip without a transition also counts as an endpoint miss. In total, 29 raters gave 529 ratings over the 197 pairs, 2.7 per pair on average. Raters agreed pairwise on 73% of the gated transition choices, 71% of the overall choices, and on 74%, 88% and 73% of statements (3), (4) and (5).

Questions as shown to raters. Verbatim from the page (single-sided wording; for pairs with both endpoints given, question 1 reads "on the start and end images" and statement 4 reads "It starts on the start image and ends on the end image."). Each question carries a one-line note in smaller type, given here in italics.

1. *A or B.* Which video reproduces the intended transition on the start image better?
   Same kind of effect, entering, spreading and clearing the same way, with similar timing.
2. *A or B.* Which video is the better effect-transfer result overall?
   Consider everything together: how well the intended transition is transferred, visual quality and aesthetics.
3. *True or False, for A and for B.* A real transition happens (not just a cut, fade or frozen image).
   Answer for each video.
4. *True or False, for A and for B.* It starts on the start image.
   Answer for each video.
5. *True or False, for A and for B.* It leaks content from the intended transition.
   A leak: people, objects or background from the intended transition's own scene show up although they are not part of the effect. Things that belong to the effect itself are not a leak.

The instruction screen adds: the videos play at their own speed and length, then wait for each other and replay together, and a shorter or less smooth video is never a reason to prefer one; the study is anonymous and only the answers are recorded.

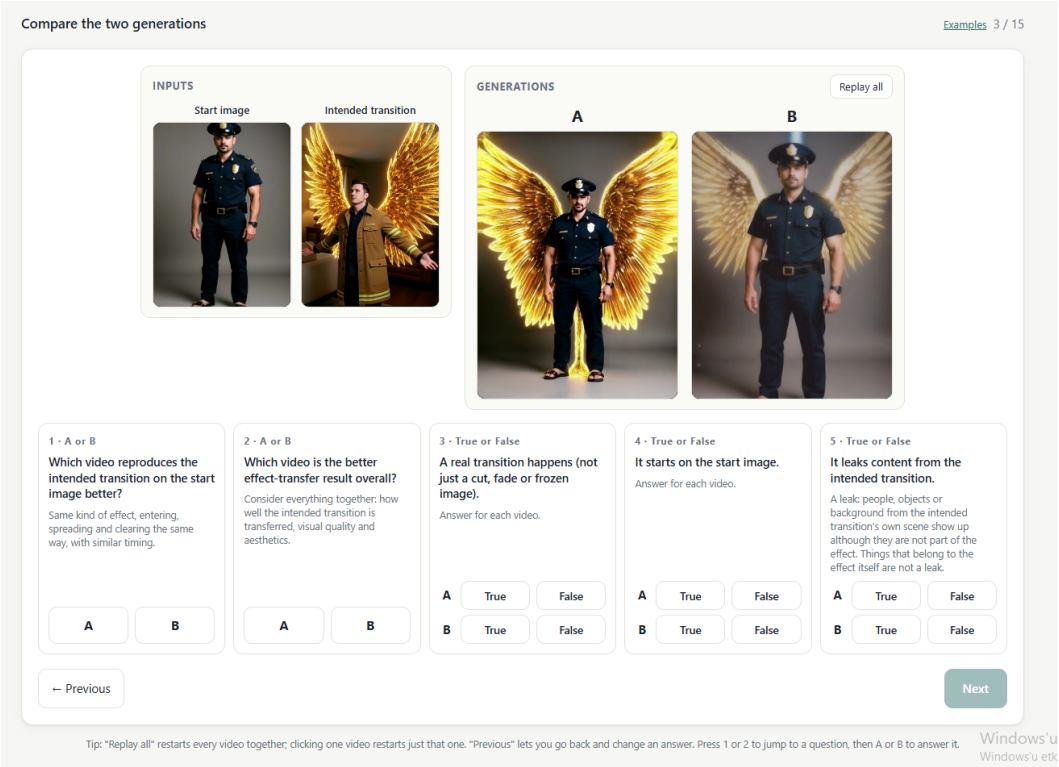

Figure 7: **The human study page** (a single-sided pair). Inputs on the left (start image, and the end image when given, then the intended transition), the two outputs on the right, and the five questions in one row: two A/B choices and three True/False statements answered for each clip. The questions are reproduced verbatim below.

VLM judge. Gemini 2.5 Pro, temperature 0, videos sampled at 8 frames per second, structured JSON output with the evidence field before the answer. Each question is asked in its own call with only the inputs it is defined on. Per pair, two forced-choice calls, (i) transition fidelity and (ii) overall, each from the given frame(s), the reference and the two outputs; each is made twice with the outputs swapped, and a pair counts as a win or a loss only when both orders agree, otherwise as a tie. Per clip, two yes/no calls: (iii) whether a transition effect occurs at all and whether content of the reference's own scenes appears, from the given frame(s), the reference and that clip alone; (iv) whether the clip's first (and last) frame is the given frame, from stills only. A clip without an effect loses both forced-choice questions by rule and is left out of the endpoint count. No system, effect or prompt is ever named to the judge. Call (i) also returns a per-output effect answer, which agrees with call (iii) on 99.8% of SEGUE clips and 98.6% of the other systems' clips. Table 5 gives the counts behind Table 4. The four instructions follow verbatim (two-endpoint variant; the one-endpoint variant drops the clauses about the given end frame).

Call (i): transition fidelity.

```
You are an expert judge of video transition effects. You receive
the GIVEN START frame as a still image, the GIVEN END frame as a
still image, a REFERENCE video, and two candidate videos, OUTPUT A
and OUTPUT B. The REFERENCE shows a transition effect between its
own two scenes. Both outputs were asked to reproduce that same
effect on the given frame(s), which show different scenes from the
reference's. No text describes the effect: the REFERENCE is the
only specification.

The outputs may come from different systems. All videos are
time-normalized to a similar duration and sampled at 8 frames per
second. Native frame rates and lengths differ, so held or repeated
frames, choppier motion or a shorter clip must never decide a
```

Table 5: **VLM judge, counts behind Table 4.** Wins / ties / losses of SEGUE on the two forced-choice questions; a tie is a pair whose choice flipped with the presentation order. Twenty-eight pairs were decided by the no-effect rule.

SEGUE vs.	Transition fidelity	Overall
Transition effect generation: both endpoints given ($n = 38$)		
refVFX (Jones et al., 2026)	34 / 4 / 0	30 / 3 / 5
Base LTX-2	28 / 6 / 4	24 / 7 / 7
Visual effect transfer: start endpoint given ($n = 183$)		
Video-As-Prompt (Bian et al., 2026)	161 / 11 / 11	149 / 14 / 20
VFXMaster (Li et al., 2025a)	117 / 37 / 29	110 / 29 / 44
refVFX (Jones et al., 2026)	143 / 20 / 20	108 / 32 / 43

```
question. Do not judge visual quality or how well the first and
last frames are kept; those are judged elsewhere.

First, for each output, state whether a transition effect occurs
at all: "yes" if something beyond a plain cut, crossfade,
dissolve, fade through black or white, blur, blend or a static
image happens; "no" otherwise.

Then answer two questions. Answer each on its own. For each, first
write one sentence of evidence citing timestamps from both
outputs, then choose A or B even when the difference is small,
then state the margin: "none", "slight" or "clear".

1. TRANSITION. Which output better reproduces the REFERENCE's
effect? An output without an effect loses this question. Otherwise
compare the effect itself: the same mechanism and material (smoke,
fire, liquid, particles, a portal, a morph, a shatter...),
entering, spreading and clearing the same way, with similar
timing. The scenes are supposed to differ: different people,
objects or backgrounds are not errors.

2. REFERENCE CONTENT. Which output shows LESS content that belongs
to the REFERENCE's own scenes: its people, objects, backgrounds,
the look of its scenes? The given frame(s) show what the output's
scene should contain; anything from the reference's scenes is
leakage. Material that is part of the effect itself (the smoke,
the fire, the particles) is not leakage. If neither output shows
any, choose either and set the margin to "none".

Return only a JSON object:
{"occurrence": {"A": "yes", "B": "yes"},
 "transition": {"evidence": "...", "choice": "A", "margin":
"clear"},
 "reference_content": {"evidence": "...", "choice": "A", "margin":
"none"}}
```

Call (ii): overall.

```
You are an expert judge of video transition effects. You receive
the GIVEN START frame as a still image, the GIVEN END frame as a
still image, a REFERENCE video, and two candidate videos, OUTPUT A
and OUTPUT B. The REFERENCE shows a transition effect between its
own two scenes. Both outputs were asked to reproduce that same
effect on the given frame(s), which show different scenes from the
reference's. No text describes the effect: the REFERENCE is the
only specification.

The outputs may come from different systems. All videos are
time-normalized to a similar duration and sampled at 8 frames per
second. Native frame rates and lengths differ, so held or repeated
```

```
frames, choppier motion or a shorter clip must never decide the
question.

Question: OVERALL, which output is the better result for this
task, the one you would keep? Weigh what a user of this effect
would care about: that the reference's effect is performed on the
given scene, that the video stays on the given scene and its
frame(s), and that it is free of visual failures a viewer would
notice. An output with no effect at all (a cut, crossfade, fade,
blur or still image) cannot be preferred. Different people,
objects or backgrounds from the reference are not errors: the
scenes are supposed to differ.

First write one sentence of evidence citing timestamps from both
outputs, then choose A or B even when the difference is small,
then state the margin: "none", "slight" or "clear".

Return only a JSON object: {"evidence": "...", "choice": "A",
"margin": "clear"}
```

Call (iii): effect present and reference leak, per clip.

```
You are an expert judge of video transition effects. You receive
the GIVEN START frame as a still image, the GIVEN END frame as a
still image, a REFERENCE video, and one OUTPUT video. The
REFERENCE shows a transition effect between its own two scenes.
The OUTPUT was asked to reproduce that same effect on the given
frame(s), which show different scenes from the reference's. No
text describes the effect: the REFERENCE is the only
specification.

Videos are sampled at 8 frames per second. Native frame rates and
lengths differ, so held or repeated frames, choppier motion or a
shorter clip never decide an answer. Do not judge how well the
effect is reproduced, how well the given frames are kept, or
visual quality; those are judged elsewhere.

Answer two yes/no questions about the OUTPUT alone. For each,
first write one sentence of evidence citing timestamps (for
example 0:01-0:02), then answer.

1. EFFECT. Does a transition effect occur in the OUTPUT at all?
"yes" if something happens beyond a plain cut, a crossfade or
dissolve, a fade through black or white, a blur or blend between
frames, or a static image. "no" otherwise. The effect does not
need to resemble the reference's to count.

2. LEAK. Does content that belongs to the REFERENCE's own scenes
appear in the OUTPUT: a recognizable person, object, background or
scene of the reference? The given frame(s) show what the OUTPUT's
scene should contain. Material that is part of the effect itself
(smoke, fire, liquid, particles, light) is not leakage. A color
cast or a vague shape is not leakage. The subject copying the pose
or action of the reference's subject is not leakage. If "yes",
name what leaked in the evidence.

Return only a JSON object:
{"effect": {"evidence": "...", "answer": "yes"},
 "leak": {"evidence": "...", "answer": "no"}}
```

Call (iv): endpoints kept, per clip, stills only.

```
You are an expert judge. You compare still frames only.

You receive the GIVEN START frame and the GIVEN END frame, then
the FIRST frame of the OUTPUT and the LAST frame of the OUTPUT.
The output was supposed to begin on the GIVEN START and end on the
```

```
GIVEN END.

Question: does the output's first frame show the GIVEN START, and
does the output's last frame show the GIVEN END? Judge only what
is visible in these stills: the same scene, the same subject and
identity, the same camera framing, the same pose. A frame that
shows all of these still counts as the given frame ("yes") even if
it is softer, blurrier, darker, color-shifted or partly warped, as
long as the scene, subject and pose remain recognizable. Answer
"no" when the frame shows a different scene, a different subject
or identity, a clearly different framing or pose, or is degraded
so far that the given frame is no longer recognizable. Ignore
anything you might guess about the video in between.

For each frame, first write one sentence naming what is the same
and what differs, then answer "yes" or "no".

Return only a JSON object: {"start": {"evidence": "...", "answer":
"yes"}, "end": {"evidence": "...", "answer": "yes"}}
```

E ADDITIONAL RESULTS

Table 6 gives the ablations on the seen and held-out instance tiers.

Table 6: **Ablations on the seen and held-out instance tiers**, both tasks included. N, E, and $|\Delta|$ as in Table 3.

	Transition fidelity ↑			Motion fid. (×100) ↑			Action KL ↓			Ref sim. (×100) ↑			Seam-free % ↑		
	N	E	$\|\Delta\|$	N	E	$\|\Delta\|$	N	E	$\|\Delta\|$	N	E	$\|\Delta\|$	N	E	$\|\Delta\|$
Seen															
Plain LoRA	85.2	92.8	7.6	**17.8**	**15.4**	2.4	4.25	4.23	**0.02**	95.9	96.0	0.1	94.2	94.2	**0.0**
SEGUE w/o NRG	92.1	95.5	3.4	13.3	13.6	**0.3**	4.35	4.31	0.04	96.1	96.1	**0.0**	96.2	96.2	**0.0**
SEGUE	**96.1**	**96.7**	**0.6**	14.2	14.5	**0.3**	**3.81**	**3.84**	0.03	**96.3**	**96.4**	0.1	**98.1**	**100.0**	1.9
Held-out instance															
Plain LoRA	72.6	84.3	11.7	15.3	14.4	0.9	5.52	5.06	0.46	95.5	95.9	0.4	**97.8**	**97.4**	**0.4**
SEGUE w/o NRG	82.5	87.6	5.1	14.8	14.7	**0.1**	5.02	4.74	0.28	95.9	96.1	0.2	95.6	94.2	1.4
SEGUE	**87.9**	**89.9**	**2.0**	**16.3**	**16.6**	0.3	**4.58**	**4.47**	**0.11**	**96.1**	**96.2**	**0.1**	92.3	91.2	1.1

Table 7 sweeps the guidance weight w of SEGUE on each tier. Transition fidelity rises with w on every tier; on the held-out instance and held-out class tiers the seam-free rate gives up a few points at $w = 6$. We use $w = 6$ throughout.

Table 7: **Guidance weight w by tier.** $w = 1$ is SEGUE w/o NRG, $w = 6$ is SEGUE.

Tier		Transition fidelity ↑	Motion fid. ↑ (×100)	Action KL ↓	Ref sim. ↑ (×100)	Seam-free % ↑
Seen	$w = 1$	92.1	13.3	4.35	96.1	96.2
	$w = 1.5$	92.8	13.9	4.09	96.2	**98.1**
	$w = 3$	94.1	13.8	4.09	96.3	**98.1**
	$w = 6$	**96.1**	**14.2**	**3.81**	**96.3**	**98.1**
Held-out instance	$w = 1$	82.5	14.8	5.02	95.9	**95.6**
	$w = 1.5$	84.0	14.7	4.89	96.0	95.3
	$w = 3$	86.5	15.5	4.80	96.1	94.5
	$w = 6$	**87.9**	**16.3**	**4.58**	**96.1**	92.3
Held-out class	$w = 1$	87.5	18.8	4.61	96.2	97.1
	$w = 1.5$	89.7	19.2	4.54	96.3	97.1
	$w = 3$	92.3	20.1	4.34	96.4	**97.3**
	$w = 6$	**93.4**	**21.0**	**4.23**	**96.4**	94.1

Figure 8 compares the systems of Table 3 with the start endpoint given, in the layout of Figure 6. Figures 9 to 12 show more results of SEGUE and more comparisons, in the layouts of Figures 1 and 6.

The supplementary material shows videos with qualitative examples. Refer to those for a better grasp of the results of SEGUE and prior art. Open the static HTML file (index.html) using a modern web browser that supports HTML5 video rendering. While Google Chrome and Mozilla Firefox are recommended, please note that native video decoding often depends on OS-specific media frameworks and codecs. If playback issues arise, verify your browser's core video decoding capabilities. For advanced users troubleshooting performance or rendering blocks, consider disabling browser extensions related to video management, ad-blocking, or memory optimization (such as Chrome Memory Saver or Firefox Efficiency Mode).

Negative reference. Section 3.3 contrasts SEGUE's dissolve null with an empty null that simply drops the reference from the null pass. Figure 13 makes the difference concrete on three examples: with the empty null the guidance amplifies the reference's own scenes along with its effect, so the ground, the running scene, and the fire hydrant of the reference leak into the output. With the dissolve null the start clip's scene is kept and only the effect is amplified.

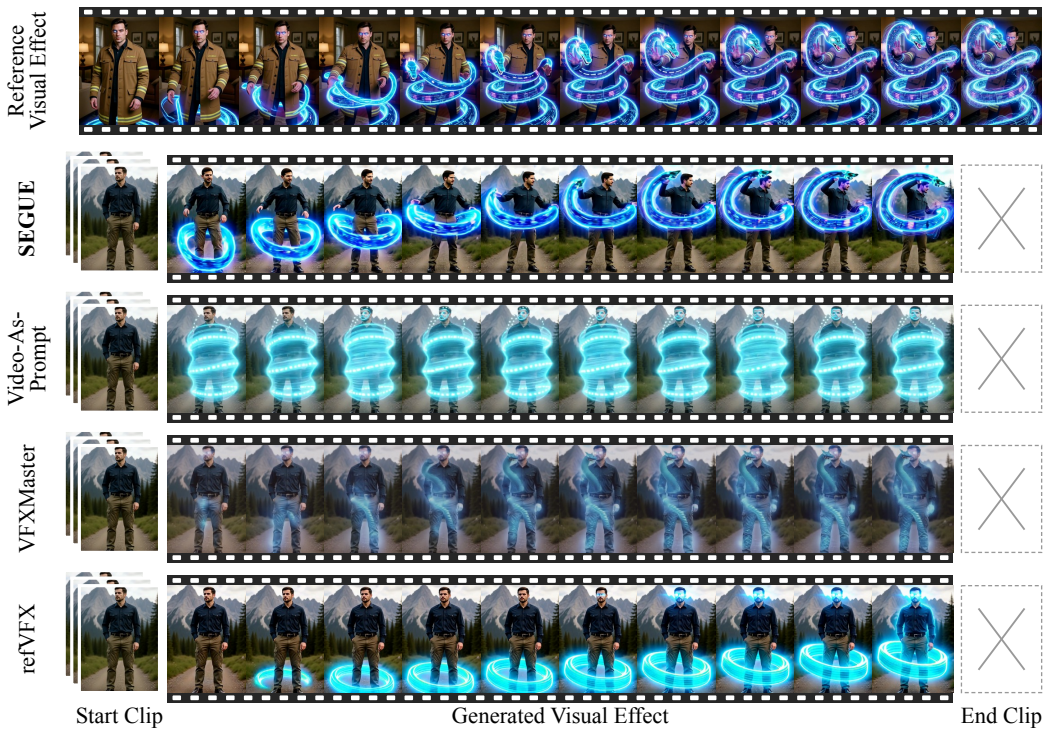

Figure 8: **Qualitative comparison**, start endpoint given. Top: the reference; below: one system per row, frames in time order.

Failure cases. Figure 14 shows the two failure modes of Section 5 on three samples: in the first two, scenes of the reference leak into the output; in the third, whose endpoint shows intense motion, the endpoints are not preserved.



Figure 9: **More results of SEGUE**, both endpoints given.

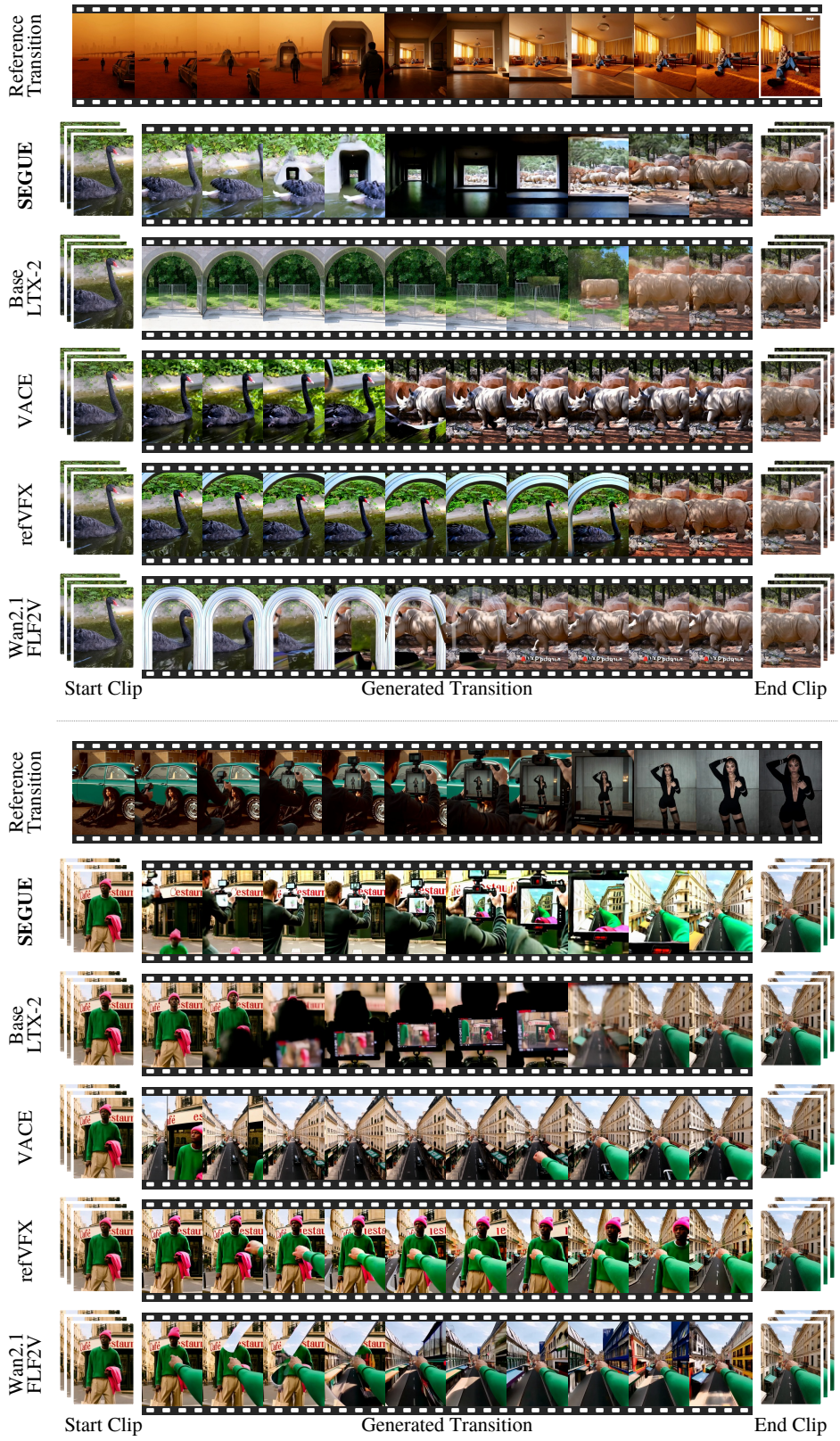

Figure 10: **More comparisons**, held-out class tier, both endpoints given, on samples where the VLM judge preferred SEGUE.



Figure 11: **More results of SEGUE**, start endpoint given.



Figure 12: **More comparisons**, held-out class tier, start endpoint given, on samples where the VLM judge preferred SEGUE.



Figure 13: **Negative reference: empty null vs. dissolve null.** Each block shows the reference clip (top strip) and SEGUE run twice with the *same* reference, endpoints, neutral prompt, seed 42 and guidance $w = 6$. Only the null reference changes. With the *empty* null (the reference dropped from the null pass) the guidance amplifies the reference's own scenes along with its effect, so reference content leaks into the output: the ground (top block), the running scene (middle) and the fire hydrant (bottom). With SEGUE's *dissolve* null (a dissolve of the reference's own endpoints) only the effect is amplified and the start clip's scene is kept.

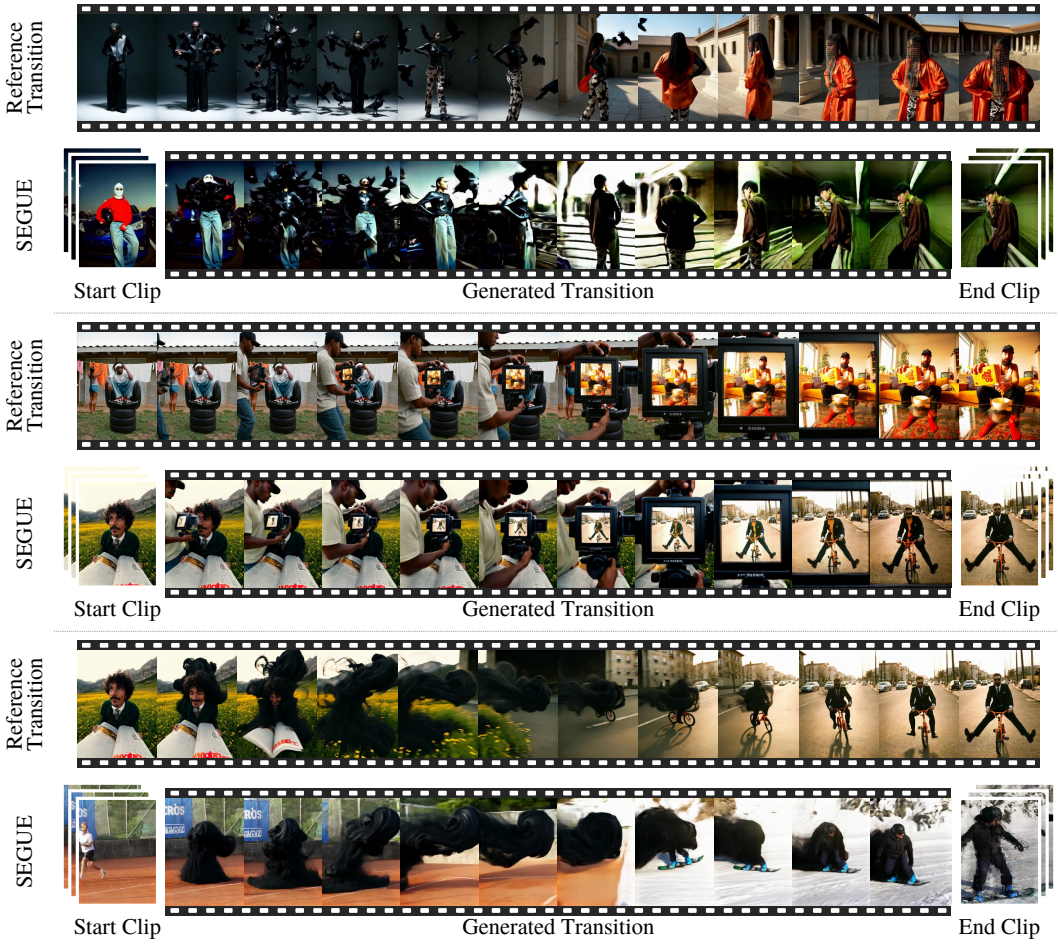

Figure 14: **Failure cases of SEGUE**, both endpoints given. Top two rows: scenes of the reference leak into the output. Bottom row: the appearance of the endpoints is not preserved through the transition.

F EVALUATION DATASET

Table 8: **Evaluation dataset by tier.** Endpoints are shared across tiers, so the total counts distinct endpoints.

			Samples	
Tier	Effect types	Distinct endpoints	Both endpoints	Start endpoint
Seen	13	26	4	22
Held-out instance	25	82	32	105
Held-out class	51	150	38	183
Total	76	204	74	310

Tiers. The evaluation dataset has 384 samples in three tiers, defined by how the reference clip relates to the training set of SEGUE. In the *seen* tier the reference clip itself was trained on. In the *held-out instance* tier the reference clip was not, but other clips of its effect were. In the *held-out class* tier the whole effect was held out: none of its 51 effects, 68 reference clips, or 150 endpoints appears in the training set. The held-out class tier is the evaluation dataset of Section 4.1; the other two tiers serve the appendix tables. Table 8 lists the three tiers.

Samples. A sample is a reference clip together with the target endpoints: both endpoints for two-sided TEG, or the start alone for single-sided visual effect transfer. The endpoints are of three kinds, so that the same reference is tested on scenes at three distances from it: a scene from another clip of the reference's own effect, never the reference clip itself (124 samples); a scene from a clip of another effect of our sources (130 samples); and a clip from DAVIS or HumanVid that no effect was ever applied to (130 samples). The seen tier uses only the first kind. Endpoints are shared across tiers, so the 204 endpoints of Table 8 are distinct.

Held-out class tier. Its 51 effects are the 17 Higgsfield transitions held out from training, each with two reference clips and six to eight samples, and the 34 EffectData effects, each with one reference clip and three samples, one of each endpoint kind. The EffectData effects are chosen as 17 couples that share a start subject, so the same subject appears under two different effects, and their endpoints are 51 EffectData subjects and 17 HumanVid clips. EffectData samples are 81 frames long, the native length of its clips; all other samples are 121 frames.

Generation. Every sample is run with two seeds. Each of our variants generates all 384 samples with the neutral prompt, 768 videos, and once more with the effect prompt for the E and $|\Delta|$ columns of Tables 3 and 6. Previous approaches generate the 183 start-only samples of the held-out class tier ($n = 366$ per system), and the adapted baselines its 38 two-endpoint samples ($n = 76$ per system).

G DISTANCE FROM LINEAR INTERPOLATION

Section 3.3 states that the base model, given two endpoints and a prompt that describes only the two scenes, mostly produces a cut or a dissolve. In pixel space both belong to the linear interpolation family ℓ,

$$V_\tau = \big(1 - \alpha(\tau)\big)\, a + \alpha(\tau)\, b, \qquad \text{for a monotonic schedule } \alpha(\tau) \in [0, 1], \tag{9}$$

where $\tau \in [0, 1]$ is normalized time, a is the last frame of the start, and b the first frame of the end. A member of ℓ only blends the two given frames: it keeps the scenes of both endpoints and adds no motion and no material of its own; the dissolve of Eq. (6) is the member with $\alpha(\tau) = \tau$. Table 9 measures how far videos sit from ℓ: the outputs of the base LTX-2 sit less than half as far from ℓ as real transitions do.

Table 9: **Distance from linear interpolation.** Mean $\pm$ s.d. over n clips.

	n	Distance from ℓ
Real transitions	73	0.54 ± 0.15
Base LTX-2 transitions	90	0.24 ± 0.11