REVERIS: ADAPTIVE REFERENCE CONDITIONING FOR FEW-STEP VISUAL EFFECT TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in visual generative models have expanded digital content creation, where creators seek to reuse distinctive effects across subjects and scenes. Reference videos provide concrete demonstrations of transformations that are difficult to specify through text alone, but also entangle the intended effect with reference-specific content. Faithful visual effect transfer therefore requires capturing the relevant change evidence while adapting its influence to the target. We introduce **Reveris**, a four-stage training framework spanning reference adaptation, modulation calibration, multi-source guidance, and few-step distillation. Within a shared image-to-video backbone, we combine clean-reference temporal variation with caption-conditioned content matching and target-conditioned modulation to adapt reference evidence to new content. We further coordinate target and reference conditions through multi-source classifier-free guidance, and distill this guided behavior through distribution matching into a four-step generator. We also construct a curated dataset of 49,194 videos spanning 350 effect types in 13 families to support learning shared effect behavior across diverse subjects and scenes. Experiments demonstrate strong overall transfer performance, with consistent improvements in effect fidelity, source preservation, and visual quality. Our project demo is available at `https://reveris123.github.io/reveris/`.

1 INTRODUCTION

Rapid advances in visual generative models have expanded the possibilities of digital content creation, enabling high-quality image and video synthesis from text and visual inputs. In digital media production, creators seek control not only over what a video depicts, but also over how its subjects and scenes transform (Li et al., 2026a). Making distinctive visual effects reusable across subjects and scenes allows creators to carry a desired visual concept into new compositions without being tied to the content of the original example.

Yet specifying an effect requires more than naming its category: its spatial patterns, intermediate appearances, and temporal progression can be difficult to convey precisely through text alone (Li et al., 2026a). A reference video provides a concrete demonstration of these properties, allowing creators to communicate the desired transformation visually rather than relying solely on a verbal description (Bian et al., 2026). *Visual effect transfer* uses such a demonstration and its caption, together with a user-provided source image and target prompt, to generate a video that applies the demonstrated effect to the source content. For example, a liquid coating spreading over a person can guide the same effect on a cup (Figure 1). Successful transfer must preserve the effect's characteristic behavior while adapting its realization to the new subject and scene.

Recent methods such as VAP, VFXMaster, and RefVFX have established reference-conditioned interfaces for visual effect transfer (Bian et al., 2026; Li et al., 2026a; Jones et al., 2026). However, a reference depicts the intended transformation together with a particular subject, background, and potentially unrelated motion. When the target differs in shape and layout, matching the reference's appearance or reproducing its motion is insufficient: the effect's characteristic evolution must be preserved, but its spatial realization must adapt to the target. This adaptation must allow the changes required by the effect while preserving unrelated source content. Our comparisons illustrate the remaining difficulty: a rising fire spiral can become a vertical flame, and a surface material conversion can become a loose covering (Figure 4). Such outputs retain a broad resemblance to the effect but

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

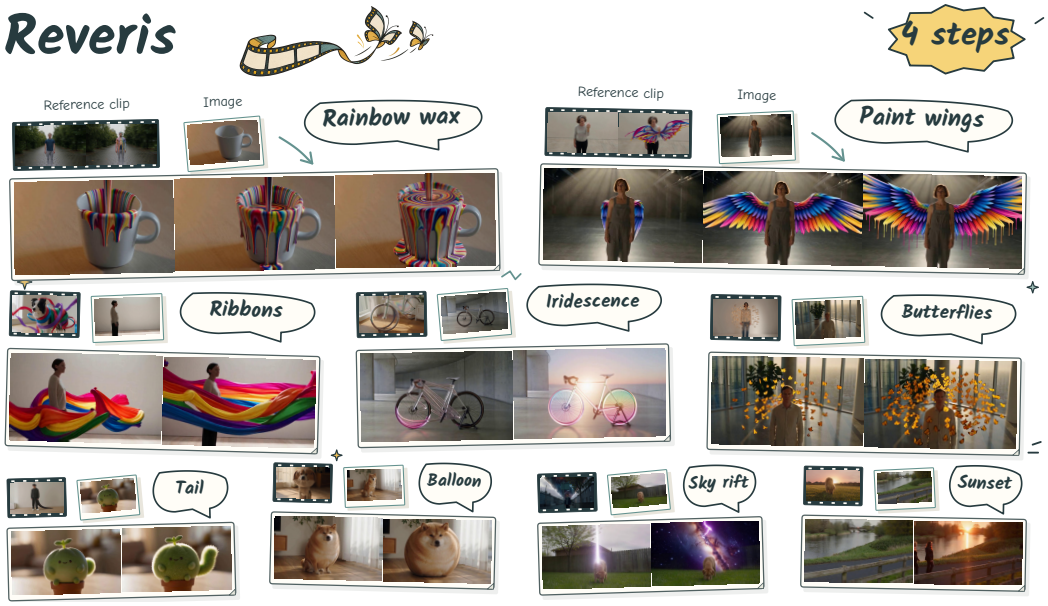

Figure 1: Reveris transfers visual effects in four denoising steps.

miss its defining dynamics or its interaction with the target's structure. The challenge is therefore not simply to inject reference features, but to determine which evidence captures the intended visual change and how its influence should depend on the target.

To this end, we introduce **Reveris**, a four-stage framework that learns adaptive reference conditioning within a shared image-to-video backbone and distills the resulting guided teacher into a four-step generator. We first fine-tune interaction self-attention to establish reference–target conditioning (Stage I). Since effects unfold through changes in appearance and geometry, temporal differences in clean reference latents provide an observable cue to their evolution. We therefore augment caption-conditioned content matching with a soft temporal-variation prior, allowing reference attention to draw on both observed change and semantic context. The influence of this cue, however, should depend on how the effect is to be realized on the target. Accordingly, we introduce adaptive reference modulation: learned signed amplitudes let attention heads favor or downweight stronger variation, while gates conditioned on the source image and target prompt scale this bias for each target. We then calibrate these new modulation parameters with the adapted backbone frozen (Stage II).

Moreover, This target-aware use of reference evidence calls for balancing target intent with the reference's semantic and visual guidance during sampling. We therefore use multi-source classifier-free guidance (CFG) (Ho & Salimans, 2022; Zhong et al., 2025), with separate scales for the incremental contributions of target text, reference text, and reference visuals. To support this control, we jointly refine reference attention and modulation under hierarchical condition dropout (Stage III). However, multi-source guidance requires four conditioned predictions at every denoising step, substantially multiplying the cost of generation. We therefore distill the combined guidance behavior into a four-step student using distribution matching distillation (DMD) (Yin et al., 2024b;a), preserving the coordinated control of multiple conditions while eliminating repeated guidance evaluations at inference (Stage IV). The student retains all input conditions and adaptive modulation while requiring only one fully conditioned prediction per step.

Learning this reference-conditioned control requires examples of the same effect across different subjects and scenes. Such cross-content supervision encourages the model to preserve shared effect behavior while adapting its realization to the target content. We therefore construct a curated dataset of 49,194 target videos spanning 350 effect types in 13 families. The dataset provides cross-content reference–target pairs with verified effect realizations and observation-based captions.

On 2,004 test cases, four-step Reveris achieves the highest mean effect-fidelity, source-preservation, and visual-quality scores among the compared methods under direct vision–language assessment.

Our contributions are summarized as follows:

- We develop Reveris, a four-stage training framework spanning reference adaptation, modulation calibration, multi-source guidance, and few-step distillation, enabling faithful and efficient transfer of demonstrated effects to new content.
- We introduce adaptive reference modulation that combines clean-reference temporal variation with content-based matching, using learned signed amplitudes and target-conditioned gates to control the cue's influence.
- We construct a curated dataset of 49,194 target videos spanning 350 effect types, with observation-based captions, verified effect realization, and within-effect pairing to supervise transfer across subjects and scenes.

2 Related Work

Visual effect generation and transfer. VFX Creator (Liu et al., 2025) combines LoRA adaptation with spatial and temporal controls; Omni-Effects (Mao et al., 2026) composes effects through LoRA experts and spatially aware prompts. For reference-based transfer, VAP (Bian et al., 2026) couples a trainable reference expert to a frozen backbone through joint attention. VFXMaster (Li et al., 2026a) uses an in-context attention mask and supports concept-token adaptation for unseen effects. RefVFX (Jones et al., 2026) jointly conditions on reference and input latents for image-to-video and video-to-video transfer. EffectMaker (Yang et al., 2026) adds effect reasoning, while P-Flow (Zhao & Shou, 2026) refines prompts using a vision–language model and a reference-derived noise prior. Their datasets use specialist generators (Liu et al., 2025; Bian et al., 2026), image editing followed by first–last-frame synthesis (Mao et al., 2026; Yang et al., 2026), or mixed pipelines (Jones et al., 2026). We adopt within-effect pairing (Li et al., 2026a; Jones et al., 2026), with observation-based captions and separate effect and quality checks. Reveris connects target-dependent reference modulation with few-step generation, addressing reference use and inference cost together.

Motion transfer and reference conditioning. LMP (Chen et al., 2025) selects foreground reference tokens through text–video attention and reweights their keys. FlexiAct (Zhang et al., 2025b) combines an image-conditioned adapter with per-reference embeddings and timestep-dependent attention. DeT (Shi et al., 2025) learns motion through shared temporal kernels and point-tracking supervision. For image conditioning, OminiControl (Tan et al., 2025) uses an attention bias to adjust conditioning strength at inference. Reveris augments video-reference attention with a temporal-variation prior and predicts its modulation from the source image and target text.

Few-step generation and guidance distillation. DMD (Yin et al., 2024b) matches output distributions through score estimates; DMD2 (Yin et al., 2024a) stabilizes training and extends it to few-step sampling. CausVid (Yin et al., 2025) distills a bidirectional video teacher into a causal generator; Self Forcing (Huang et al., 2025) trains on autoregressive self-rollouts; rCM (Zheng et al., 2026a) combines consistency with score regularization. For controllable generation, FlashMotion (Li et al., 2026b) adapts trajectory-control modules to a distilled backbone, while FADA (Zhong et al., 2025) distills audio and reference guidance through learnable guidance tokens. Reveris distills the reference-modulated teacher's target-text, reference-text, and reference-video guidance at fixed scales into four fully conditioned predictions, reducing denoising steps and CFG cost.

3 Dataset Construction

Learning reference-conditioned transfer requires examples of the same effect across different subjects and scenes. We generate and verify multiple realizations of each effect, then pair accepted videos to create transfer training examples (Figure 2).

Generating effect videos. We organize effects by their visual transformation, including material changes, shape deformation, particle phenomena, and environmental changes. Each effect has a recipe specifying the affected region, required initial state, temporal progression, visible intermediate states, and final outcome. For each recipe, we use a vision–language model (VLM) to describe concrete scenes with people, animals, objects, vehicles, or environments, varying appearance and background. We generate a source clip and retain its first frame when the required subject or surface

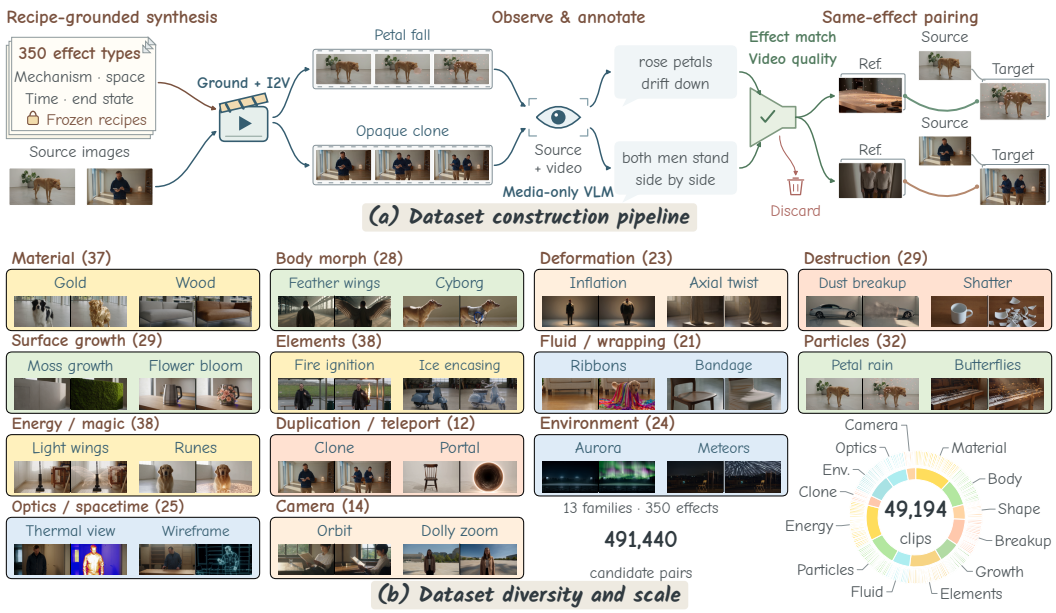

Figure 2: Dataset construction and effect diversity.

is clearly visible and the effect has not yet occurred. The VLM then identifies the entities and spatial locations in this image to which the recipe applies. We combine this information with the recipe to form an image-to-video prompt specifying the transformation, content to preserve, and continuous progression from the initial state. We animate the source image with Seedance 2.0, Seedance 2.5, Kling, and Veo 3.1 to obtain candidate effect videos.[1]

Captioning and filtering generated videos. Generation may deviate from the recipe, so we annotate what each video actually depicts. Given the source image and generated video, the VLM records the subject, affected region, onset and progression of the change, intermediate states, final appearance, and camera motion. These observations provide a separate training caption for each video. We then check whether the observed transformation realizes the intended effect and assess source correspondence, shot continuity, structural integrity, and unrelated changes. A candidate must satisfy both effect and quality checks: missing or incorrect effects, shot changes, severe artifacts, and dominant unrelated transformations cause rejection. Minor differences in timing, color, or intensity are accepted when the defining effect remains visible and the video is usable. Accepted videos retain their original effect labels, forming pools of captioned examples with varied visual realizations.

Forming transfer training samples. For each retained target video, we sample up to ten other videos without replacement from the same effect pool and split, excluding the target itself. During training, we sample one reference and combine it with the source image used to generate the target. The target and reference each retain their own observed caption; these paired visual and textual conditions supervise generation of the target video. Reference and target share an effect category while depicting different subjects or scenes, with no required framewise correspondence. The dataset contains 49,194 target videos across 350 effect types in 13 families, yielding 491,440 candidate reference–target pairs.

4 METHOD

Given a source image I, a reference video R, target text p, and a reference caption p_r, we generate a video Y that realizes the demonstrated effect on the source content. Let $z_0 = \mathcal{E}(Y)$ and $r = \mathcal{E}(R)$ be VAE latents. We corrupt only the target as $z_t = (1-t)z_0 + t\epsilon$, with $t \in [0,1]$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$; the reference stays clean. The full condition is $c = (I, p, p_r, \mathcal{R})$, where $\mathcal{R}$ groups the reference

[1]Model information: Seedance 2.0, Seedance 2.5, Kling, and Veo 3.1.

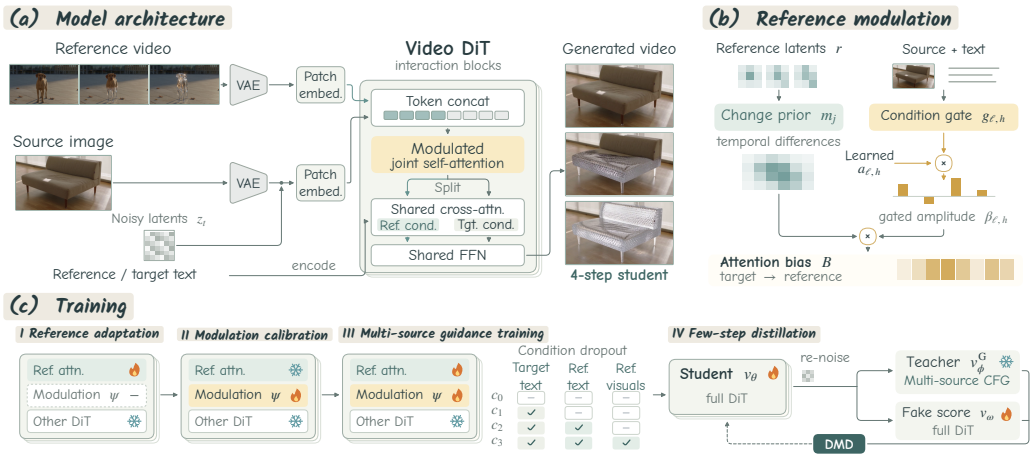

Figure 3: Reveris combines adaptive reference conditioning with guided distillation.

video and its first-frame image conditions. We train a teacher $v_\phi(z_t, t; c)$ to predict $\epsilon - z_0$ with flow matching (Lipman et al., 2023), then distill it into a four-step student.

4.1 Reference Adaptation

We adapt a pretrained Wan I2V backbone (Team Wan et al., 2025) to process target and reference token groups (Figure 3). In designated interaction blocks, the two groups are concatenated for joint self-attention, then each group cross-attends to its own image and text conditions through a shared cross-attention module. Other blocks update only target tokens, while reference tokens bypass them. In *Stage I: reference adaptation*, we fine-tune interaction self-attention to establish reference conditioning without adding parameters. Further conditioning details appear in Appendix A.

4.2 Adaptive Reference Conditioning

We augment content-based matching with reference temporal variation. Evolving effects alter local appearance and geometry, motivating attention to changing reference content. We measure this cue in the clean reference r, since differences in noisy target latents z_t also reflect diffusion noise.

For a C-channel reference latent vector $r_{\tau,u}$ at time τ and spatial location u, define

$$d_{\tau,u} = \frac{\|r_{\tau,u} - r_{\tau-1,u}\|_2}{\sqrt{C}}, \qquad m = \mathcal{N}_\kappa(\mathrm{Pool}(d)) \,. \tag{1}$$

Pooling aligns differences with reference tokens; $\mathcal{N}_\kappa$ normalizes by their per-reference mean and clips at κ. The resulting m_j records local visual change, including motion and appearance changes.

We use m_j as a soft attention bias and condition reference features on p_r through cross-attention to supply semantic context for content-based matching. To learn how each head uses this cue, we assign each interaction block ℓ and head h a trainable scalar amplitude $a_{\ell,h} \in \mathbb{R}$, initialized to zero and shared across inputs. To adapt its influence to the target, we pool projected source-image and target-text tokens into $\bar{e}$ and predict a gate:

$$g_{\ell,h}(I,p) = \mathrm{sigmoid}([W_\ell \bar{e} + b_\ell]_h) \,, \qquad \beta_{\ell,h} = a_{\ell,h}\, g_{\ell,h}(I,p). \tag{2}$$

The amplitude determines the sign of the bias; the gate scales its magnitude for each input.

For joint projections Q, K, V, we bias attention from target queries to reference keys:

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}} + B\right) V, \quad B_{ij} = \begin{cases} \beta_{\ell,h} m_j, & i \in \mathcal{I}_{\mathrm{tar}},\ j \in \mathcal{I}_{\mathrm{ref}}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where d_h is the head dimension and $\mathcal{I}_{\mathrm{tar}}, \mathcal{I}_{\mathrm{ref}}$ index target and reference tokens. A positive $\beta_{\ell,h}$ favors reference keys with greater temporal variation; a negative value reverses this preference. The bias combines with content-based similarities to determine attention.

In *Stage II: modulation calibration*, we freeze the adapted backbone and train only the new parameters $\psi = \{a_\ell, W_\ell, b_\ell\}_\ell$ using the same flow-matching objective.

4.3 Multi-Source Guidance

Reference text and video provide semantic and visual controls alongside the target prompt, motivating CFG for each source. We retain the source image and introduce these controls progressively: target text establishes the desired output, followed by the reference description and then its visual realization. In *Stage III: multi-source guidance training*, we jointly update ψ and the interaction self-attention layers, keeping the remaining parameters frozen. We train these predictions with hierarchical condition dropout (Geffner et al., 2025), sampling one of four nested states per batch:

$$\begin{aligned} c_0 &= (I, \emptyset, \emptyset, \emptyset), \quad & c_1 &= (I, p, \emptyset, \emptyset), \\ c_2 &= (I, p, p_r, \emptyset), \quad & c_3 &= (I, p, p_r, \mathcal{R}). \end{aligned} \tag{4}$$

Here $\emptyset$ denotes a null condition. Writing $v_j = v_\phi(z_t, t; c_j)$, the teacher combines these predictions using multi-source guidance (Zhong et al., 2025):

$$v_\phi^{\mathrm{G}} = v_0 + s_{\mathrm{tar}}(v_1 - v_0) + s_{\mathrm{txt}}(v_2 - v_1) + s_{\mathrm{vis}}(v_3 - v_2). \tag{5}$$

Each difference measures a condition's contribution given the preceding conditions. Separate scales control textual guidance and visual-reference strength.

4.4 Few-Step Distillation of the Guided Teacher

Multi-source guidance requires four conditioned teacher predictions at every denoising step. To reduce both denoising steps and repeated guidance evaluations, we distill the guided teacher in *Stage IV*. We freeze the Stage III teacher and initialize a student v_θ and an auxiliary fake-score network v_ω from its weights. Following DMD/DMD2 (Yin et al., 2024b;a), we match the student's generated distribution to the guided teacher. The teacher uses fixed guidance scales; both trainable networks receive c_3. For a clean student prediction $x = z_{t_k} - t_k v_\theta(z_{t_k}, t_k; c_3)$, form $y_\sigma = (1 - \sigma)x + \sigma\eta$ with fresh Gaussian noise η and $\sigma \in (0, 1)$. The teacher and fake-score networks estimate

$$d_{\mathrm{T}} = y_\sigma - \sigma v_\phi^{\mathrm{G}}(y_\sigma, \sigma), \qquad d_{\mathrm{F}} = y_\sigma - \sigma v_\omega(y_\sigma, \sigma; c_3). \tag{6}$$

With D latent elements and $\rho = D^{-1}\|x - d_{\mathrm{T}}\|_1$, the normalized DMD surrogate is

$$\mathcal{L}_{\mathrm{DMD}}(\theta) = \mathbb{E}\left[\frac{1}{2D}\,\|x - \mathrm{sg}[x - \Delta]\|_2^2\right], \qquad \Delta = \frac{d_{\mathrm{F}} - d_{\mathrm{T}}}{\rho}. \tag{7}$$

Here sg stops gradients, yielding the student gradient Δ/D with respect to x. Because d_{T} uses the combined prediction in Eq. (5), this supervision incorporates all three guidance increments. To track the changing student distribution, we alternate student updates with flow-matching updates of v_ω:

$$\mathcal{L}_{\mathrm{fake}}(\omega) = \mathbb{E}\left[\frac{1}{D}\,\|v_\omega(y_\sigma, \sigma; c_3) - (\eta - \mathrm{sg}[x])\|_2^2\right]. \tag{8}$$

For this update, y_σ is formed from detached student predictions with fresh noise. We train all DiT parameters of both networks.

During training, we run earlier denoising and re-noising steps without gradients and backpropagate through a uniformly selected step k. The four-step sampler retains all reference conditions and adaptive modulation, making one fully conditioned prediction per step.

5 Evaluation Protocol

Test cases. We evaluate on 2,004 cases spanning diverse subjects and scenes, with video content unseen during training. Each case supplies a source image, reference video, target prompt, and reference caption. To prevent leakage, test videos are excluded from all training stages, including distillation, in both target and reference roles. All methods share the test inputs and scoring protocol.

Video quality and transfer fidelity. Successful transfer must realize the demonstrated effect on the new source while maintaining video quality. We report five VBench metrics (Huang et al., 2024):

Table 1: Transfer quality on 2,004 test cases. Best and second-best displayed values are bold and underlined. Definitions are in Section 5.

		GPT-5.6 Sol (0–5)						
Method	Steps	EF↑	SP↑	TC↑	RD↑	VQ↑	EL↑	Avg. Rank↓
VAP	50	2.530	3.247	2.937	4.961	3.036	3.120	7.50
VFXMaster	50	3.020	3.499	3.123	4.951	2.932	3.581	5.67
RefVFX	50	2.735	3.375	3.100	4.973	3.034	3.443	6.33
RefVFX+CausVid	6	2.778	3.543	3.358	4.974	3.153	3.400	4.50
Reveris w/o prior	50	2.927	3.550	3.326	4.980	3.159	3.635	3.17
Reveris w/o gate	50	2.597	3.465	3.143	4.971	3.101	3.574	5.67
Reveris Teacher	50	3.073	3.628	3.407	4.977	3.249	3.734	2.17
Reveris (Ours)	4	**3.110**	**3.700**	**3.468**	**4.982**	**3.392**	**3.892**	**1.00**

	Agent (0–1)				VBench (0–1)				
Method	SF↑	SRD↑	TC↑	RL↑	SC↑	BC↑	MS↑	AQ↑	IQ↑
VAP	0.685	0.916	0.794	0.932	0.821	0.888	0.983	0.547	0.596
VFXMaster	0.808	0.937	0.859	0.930	0.861	0.901	**0.989**	0.538	0.612
RefVFX	0.745	0.930	0.845	0.929	0.845	0.898	0.988	0.556	0.639
RefVFX+CausVid	0.753	0.942	0.868	0.931	**0.868**	0.906	0.986	**0.582**	**0.693**
Reveris w/o prior	0.798	0.942	0.849	0.934	0.836	0.897	0.985	0.562	0.629
Reveris w/o gate	0.704	0.935	0.860	0.934	0.848	0.901	0.987	0.566	0.632
Reveris Teacher	0.811	0.947	0.863	**0.936**	0.850	0.904	0.987	0.559	0.627
Reveris (Ours)	**0.822**	**0.951**	**0.882**	**0.936**	0.858	**0.908**	0.988	0.566	0.637

subject consistency (SC), background consistency (BC), motion smoothness (MS), aesthetic quality (AQ), and imaging quality (IQ). These output-only measures can reward a consistent video that omits the effect, whereas a valid transformation may require appearance changes. To assess transfer directly, GPT-5.6 Sol jointly inspects the source image, sampled reference and output frames, and task descriptions. It assigns 0–5 scores for effect fidelity (EF), source preservation (SP), temporal consistency (TC), reference disentanglement (RD), visual quality (VQ), and locality of changes (EL), allowing effect-required changes while penalizing incidental reference copying.

Evidence-grounded evaluation agent. Our agent first describes the output without access to source, reference, or task information (Han et al., 2025; Zhang et al., 2025a). It then compares this observation with the reference using motion, appearance-change, subject-similarity, and leakage evidence. Uncertain judgments receive targeted review, including crops of suspected reference leakage. The agent scores semantic fidelity (SF), subject–reference disentanglement (SRD), temporal coherence (TC), and reference leakage (RL) on 0–1 scales; higher RL means less leakage.

Aggregation and inference cost. Each evaluator uses fixed prompts and inference settings across methods (Appendix B). We average judge scores equally over cases and use VBench's official aggregation. Avg. Rank averages ranks across the six direct-VLM dimensions; lower is better (Han et al., 2025). Latency covers condition encoding, denoising, and decoding; denoiser evaluations (NFE) count every CFG branch.

6 Experiments

6.1 Experimental Setup

Reveris builds on Wan2.1-I2V-14B-480P (Team Wan et al., 2025) and denotes the four-step student unless specified otherwise. *Reveris Teacher* is the 50-step Stage III model with guidance scales $(4, 5, 2)$. We compare with VAP (Bian et al., 2026), VFXMaster (Li et al., 2026a), RefVFX (Jones et al., 2026), and six-step RefVFX+CausVid (Yin et al., 2025). All methods use the same source images, reference videos, and target prompts; reference captions are supplied where supported by their native interfaces. Each output contains 49 frames at 832×480 and 16 fps. Scoring follows

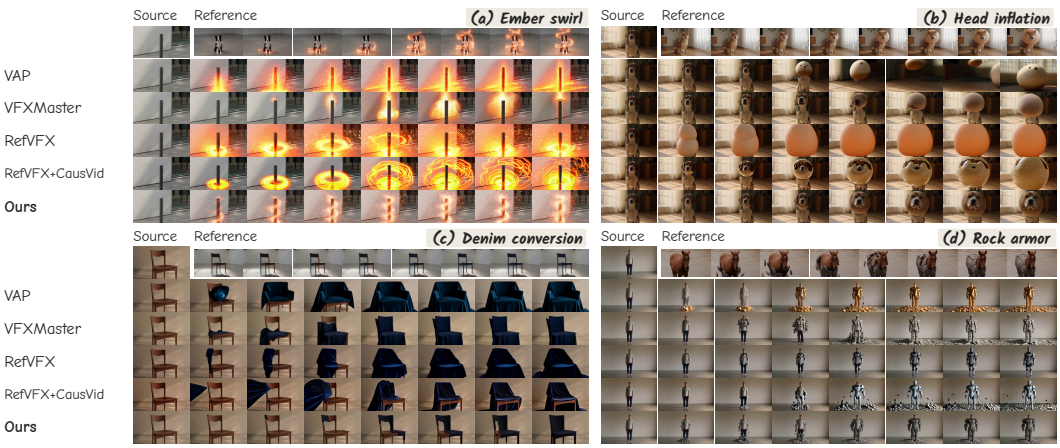

Figure 4: Qualitative comparison at matched output frames.

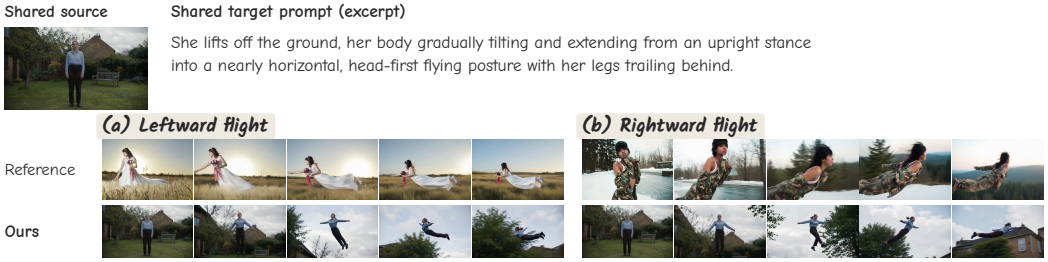

Figure 5: Reference video and caption vary jointly; source image, target text, and seed are fixed.

Section 5; RefVFX uses a public reimplementation. The appendix provides training and sampling configurations and supplementary analyses.

6.2 Quantitative Comparison

Transfer performance. Reveris ranks first across all six direct-VLM dimensions in Table 1. Relative to the strongest external baseline on each axis, it improves EF by 0.090, SP by 0.156, and VQ by 0.239. These joint gains indicate that the demonstrated transformation is reproduced while the source subject and scene are preserved. The highest agent SF, SRD, and TC scores provide complementary support for faithful transfer with coherent motion and limited reference-content leakage.

RefVFX+CausVid leads VBench SC, AQ, and IQ but trails on EF, illustrating the distinction between output quality and reference fidelity. Additional evaluation on 500 EffectData cases (Yang et al., 2026) and with alternative VLMs supports the transfer-quality trends.

Table 2: Full-video cost on one A100 80GB with CPU offloading (ten cases, three repeats). Latency includes condition encoding and decoding; NFE counts all CFG branches.

Method	Steps	NFE	Latency (s)↓
VAP	50	100	946.60
VFXMaster	50	100	812.39
RefVFX	50	150	1651.47
RefVFX+CausVid	6	18	242.49
Reveris Teacher	50	200	1636.34
Reveris (Ours)	**4**	**4**	**71.72**

Quality and efficiency after distillation. The student retains similar EF to the teacher (3.110 vs. 3.073) and improves SP and VQ by 0.072 and 0.143, respectively. Distillation reduces both sampling and guidance costs: the teacher evaluates four conditioning states at each of 50 steps, whereas the student requires four predictions in total. Full-video generation takes 71.72 s (Table 2), giving 22.8× acceleration over the teacher and 3.38× over six-step RefVFX+CausVid. The resulting model retains the guided teacher's transfer quality while achieving the lowest measured generation cost among the compared methods.

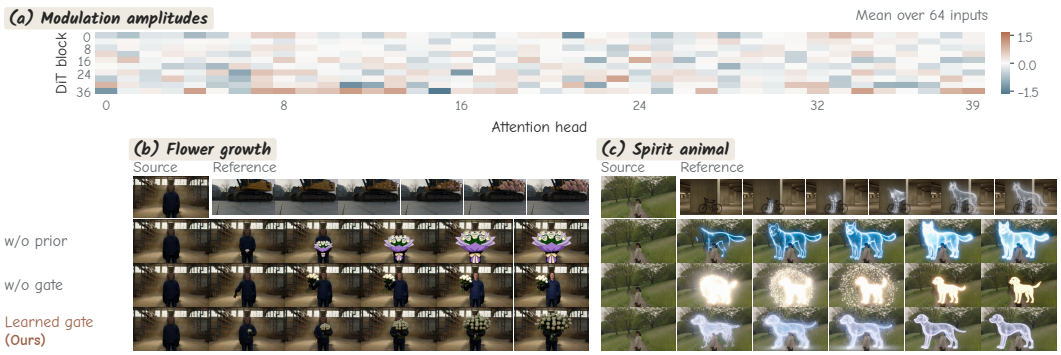

Figure 6: Reference modulation: mean effective amplitudes $\beta_{\ell,h}$ and qualitative ablations.

6.3 Qualitative Comparison

Effect structure and target fit. In Figure 4, VAP turns ember swirl into a vertical flame and RefVFX+CausVid produces horizontal rings, while Reveris follows the reference's rising spiral. For denim conversion, Reveris fits the material to the chair's surface, whereas several baselines drape cloth over it. The head-inflation example further illustrates identity preservation during deformation.

Paired reference control. In Figure 5, we fix the source image, target prompt, and seed, and change the reference video together with its matched caption. Although the target prompt leaves flight direction open, four-step Reveris follows the opposite reference directions while retaining the source subject and scene. This example shows that the paired reference steers a property unspecified by the target text, and that this responsiveness survives distillation.

6.4 Ablation Studies

Adaptive reference modulation. We compare three separately trained designs: reference interaction alone (*w/o prior*), a temporal-variation bias with learned, target-independent amplitudes (*w/o gate*), and the adaptive teacher. Table 1 shows that adaptive modulation improves EF and SP over interaction alone by 0.146 and 0.078. The ungated variant falls below interaction alone on EF (2.597 vs. 2.927). These comparisons favor target-dependent modulation of the cue.

In Figure 6, adaptive modulation spreads flowers across the subject and preserves the translucent animal's outline; the alternatives limit flower growth or miss translucency. The mean effective amplitudes contain both positive and negative values, reflecting different preferences for reference variation across heads. With teacher weights, inputs, and noise fixed, replacing the change map by its mean or shuffling its positions alters object shape and effect progression in the diagnostic examples, showing sensitivity to cue placement.

Multi-source guidance. Across five settings of the fixed teacher, $(1, 1, 1)$ retains all inputs without CFG extrapolation but gives the weakest effect fidelity. Increasing only visual guidance from $(5, 5, 2)$ to $(5, 5, 5)$ raises EF while reducing SP, TC, and VQ. Our default $(4, 5, 2)$ yields the highest mean direct-VLM TC and VQ, with aggregate performance close to $(4, 4, 2)$. Qualitative comparisons further favor the default's control over the affected region and completion of the transformation. Together, these results support balancing the three guidance sources before distillation.

7 Conclusion

We presented Reveris, a four-stage framework for adaptive reference conditioning and few-step visual effect transfer. Clean-reference temporal variation and target-conditioned attention modulation adapt demonstrated effects to new content. Multi-source guidance coordinates target and paired reference conditions; distribution matching distillation transfers this guidance to a four-step student. Supported by our curated transfer dataset, Reveris achieves strong effect fidelity and visual quality on 2,004 test cases, with $22.8\times$ faster generation than its 50-step teacher.

AI USE STATEMENT

In this work, we used generative AI tools for constructing and annotating synthetic datasets, refining research ideas and experimental designs, qualitative analysis and interpretation of results, and translation. Generative video models produced candidate training videos, while vision–language models supported scene descriptions, video captioning, filtering, and automated evaluation. Tasks involving formal mathematical proofs are not applicable to this empirical study. Additionally, we used generative AI tools for literature retrieval, manuscript drafting and editing, figure preparation, and development of supporting data-processing, evaluation, and visualization code. We have reviewed all AI-assisted work. Specifically, we checked technical descriptions and reported results against the implementation and experimental records, and verified cited references against their sources. We take responsibility for the final content of this work, including text, claims, data, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Reveris is intended to support creative visual effect production. Like other video generation methods, it could be misused to create misleading depictions or unauthorized transformations of identifiable people. Responsible use should respect privacy, obtain appropriate permission for source and reference media, and clearly identify synthetic outputs when they could be mistaken for authentic recordings. Our training corpus is synthetically generated, and evaluation relies on automated metrics and vision–language models. These data and evaluators may reflect biases inherited from their underlying models, which should be considered when interpreting results and deploying the method. Use and distribution of related resources should respect applicable licenses and usage terms.

REFERENCES

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*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ba8d1b46292c5e82cbfb3b3dc3b968af-Abstract-Conference.html.

Changgu Chen, Xiaoyan Yang, Junwei Shu, Changbo Wang, and Yang Li. LMP: Leveraging motion prior in zero-shot video generation with diffusion transformer. *arXiv preprint arXiv:2505.14167*, 2025. URL https://arxiv.org/abs/2505.14167.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteína: Scaling flow-based protein structure generative models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f4e9121ad30cd4e5528042fbfd835b3f-Paper-Conference.pdf.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-Bench: Human-aligned video generation benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18858–18868, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Han_Video-Bench_Human-Aligned_Video_Generation_Benchmark_CVPR_2025_paper.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5576. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

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*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

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*, 2026. URL https://arxiv.org/abs/2601.07833.

Fangyu Lei, Jinxiang Meng, Yiming Huang, Junjie Zhao, Yitong Zhang, Jianwen Luo, Xin Zou, Ruiyi Yang, Wenbo Shi, Yan Gao, Shizhu He, Zuo Wang, Qian Liu, Yang Wang, Ke Wang, Jun Zhao, and Kang Liu. DAComp: Benchmarking data agents across the full data intelligence lifecycle. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2512.04324.

Baolu Li, Yiming Zhang, Qinghe Wang, Liqian Ma, Xiaoyu Shi, Xintao Wang, Pengfei Wan, Zhenfei Yin, Yunzhi Zhuge, Huchuan Lu, and Xu Jia. VFXMaster: Unlocking dynamic visual effect generation via in-context learning. In *ACM SIGGRAPH 2026 Conference Papers*, pp. 1–9, 2026a. doi: 10.1145/3799902.3811208. URL https://doi.org/10.1145/3799902.3811208.

Quanhao Li, Zhen Xing, Rui Wang, Haidong Cao, Qi Dai, Daoguo Dong, and Zuxuan Wu. FlashMotion: Few-step controllable video generation with trajectory guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.

8986–8996, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_FlashMotion_Few-Step_Controllable_Video_Generation_with_Trajectory_Guidance_CVPR_2026_paper.html.

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*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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*, 2024. URL https://arxiv.org/abs/2303.05499.

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. URL https://arxiv.org/abs/2502.05979.

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. doi: 10.1609/aaai.v40i10.37737. URL https://ojs.aaai.org/index.php/AAAI/article/view/37737.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, 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.

Yusu Qian, Jiasen Lu, Tsu-Jui Fu, Xinze Wang, Chen Chen, Yinfei Yang, Wenze Hu, and Zhe Gan. GIE-Bench: Towards grounded evaluation for text-guided image editing. *arXiv preprint arXiv:2505.11493*, 2025. URL https://arxiv.org/abs/2505.11493.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45c1f6a8cbf2da59ebf2c802b4f742cd-Abstract-Conference.html.

Qingyu Shi, Jianzong Wu, Jinbin Bai, Jiangning Zhang, Lu Qi, Yunhai Tong, and Xiangtai Li. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10995–11005, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Shi_Decouple_and_Track_Benchmarking_and_Improving_Video_Diffusion_Transformers_For_ICCV_2025_paper.html.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0c38f54740062529aa4117a04b583f3c-Abstract-Conference.html.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. OminiControl: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Tan_OminiControl_Minimal_and_Universal_Control_for_Diffusion_Transformer_ICCV_2025_paper.html.

Team Wan, Ang Wang, Baole Ai, Bin Wen, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/3526_ECCV_2020_paper.php.

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, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_EffectMaker_Unifying_Reasoning_and_Generation_for_Customized_Visual_Effect_Creation_CVPR_2026_paper.html.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 47455–47487, 2024a. doi: 10.52202/079017-1505. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/54dcf25318f9de5a7a01f0a4125c541e-Abstract-Conference.html.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.html.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_From_Slow_Bidirectional_to_Fast_Autoregressive_Video_Diffusion_Models_CVPR_2025_paper.html.

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation Agent: Efficient and promptable evaluation framework for visual generative models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7561–7582. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-long.374. URL https://aclanthology.org/2025.acl-long.374/.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Shiyi Zhang, Junhao Zhuang, Zhaoyang Zhang, Ying Shan, and Yansong Tang. FlexiAct: Towards flexible action control in heterogeneous scenarios. In *ACM SIGGRAPH 2025 Conference Papers*, pp. 1–11, 2025b. doi: 10.1145/3721238.3730683. URL https://doi.org/10.1145/3721238.3730683.

Rui Zhao and Mike Zheng Shou. P-Flow: Prompting visual effects generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9149–9160, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhao_P-Flow_Prompting_Visual_Effects_Generation_CVPR_2026_paper.html.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0534abc9e6db91683d82186ef0d68202-Abstract-Conference.html.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cdb347d7516d52a9280ea9d6708911f6-Abstract-Conference.html.

Tianyun Zhong, Chao Liang, Jianwen Jiang, Gaojie Lin, Jiaqi Yang, and Zhou Zhao. FADA: Fast diffusion avatar synthesis with mixed-supervised multi-CFG distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3101–3110, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhong_FADA_Fast_Diffusion_Avatar_Synthesis_with_Mixed-Supervised_Multi-CFG_Distillation_CVPR_2025_paper.html.

A Additional Method Details

This appendix specifies the reference-conditioning operations, teacher training, and guided distillation used in Section 4.

A.1 Reference Conditioning

Token groups and positional encoding. Both token groups use the pretrained I2V conditioning interface: video latents are accompanied by first-frame latent conditions and image embeddings, together with the corresponding text. The target receives the source image and target text; the reference receives its own first frame and caption. The reference tokens use a zero noise timestep. Following VAP (Bian et al., 2026), its temporal token coordinates occupy $-T'_r, \dots, -1$, preceding target coordinates $0, \dots, T'_y - 1$, where T'_r and T'_y are token-grid lengths. Spatial coordinates are unchanged. Joint self-attention concatenates the two token groups along the sequence dimension and retains the backbone's query/key normalization and rotary position encoding. For cross-attention and feed-forward updates, the two groups are stacked along the batch dimension and processed by the same modules, with each batch entry retaining its own conditions. Our DiT has 40 blocks with 40 heads, with reference interactions enabled at zero-based block indices $\{0, 4, \dots, 36\}$.

Variation prior. For $r \in \mathbb{R}^{C \times T_r \times H_r \times W_r}$, Eq. (1) is evaluated at $\tau = 2, \dots, T_r$. We copy the first available difference to $d_{1,u}$; a single-frame reference has zero variation. Average pooling uses the DiT patch size and stride, producing $\widetilde{d}_j$ for each reference token; our model uses $C = 16$ and a $(1, 2, 2)$ temporal–spatial patch size. The normalization operator in the main text is

$$m_j = \begin{cases} \min\left(\kappa, \widetilde{d}_j / \bar{d}\right), & \bar{d} > \varepsilon_0, \\ 0, & \bar{d} \le \varepsilon_0, \end{cases} \qquad \bar{d} = \frac{1}{N_r} \sum_{j=1}^{N_r} \widetilde{d}_j, \tag{A.1}$$

where N_r is the number of reference tokens, $\varepsilon_0 = 10^{-6}$, and $\kappa = 3$. The threshold handles references with negligible latent variation, while clipping limits extreme differences. The prior is computed from the clean video latent channels before patch embedding; auxiliary first-frame conditioning channels do not enter this calculation. Differences computed from noisy target latents would also reflect differences in the sampled noise, which can dominate at high noise levels. Using the clean reference provides an observed temporal cue throughout denoising.

Attention bias and conditioning. The trainable scalar $a_{\ell,h}$ sets the sign and baseline magnitude of the bias and is shared across inputs; the target-conditioned gate scales its magnitude for each input. All target queries in one head share this reference-key bias. The gate receives pooled source-image and target-text tokens, while reference variation enters through m. For fixed conditions, gates remain constant across denoising steps. Reference text supplies semantic context through cross-attention to the reference token group. These caption-conditioned features participate in subsequent interaction blocks and provide semantic context for content-based matching. This feature-conditioning path leaves the calculation of m unchanged.

A.2 Teacher Training and Conditioning States

Supervised objective. Stages I–III minimize

$$\mathcal{L}_{\mathrm{FM}}(\phi) = \mathbb{E}_{z_0, \epsilon, t, c}\left[\frac{1}{D} \left\| v_\phi(z_t, t; c) - (\epsilon - z_0) \right\|_2^2\right], \tag{A.2}$$

where D is the number of target latent elements and $z_t = (1 - t)z_0 + t\epsilon$. Stages I and II use the full condition. Stage III uses the same objective with a sampled state from Eq. (4). The training variable t denotes the noise proportion; the network receives the corresponding timestep scaled by 1,000. Stage I reuses the pretrained modules for both token groups and introduces no new parameters; only the designated self-attention weights are updated. Stage II introduces the amplitudes and gate projections, initializing their parameters to zero so that the added attention bias starts at zero and the gates start at 0.5. With ten interaction blocks, 40 heads per block, and a hidden dimension of 5,120, the amplitudes and linear gate projections introduce $10 \times (40 + 5{,}120 \times 40 + 40) = 2{,}048{,}800$ parameters in total. The teacher and student share this architecture.

Hierarchical condition dropout. Three independently droppable condition groups admit $2^3 = 8$ subsets. Our guidance rule instead uses an ordered path through these subsets: target text specifies the desired output, reference text adds the demonstration's description, and reference visuals supply its concrete realization. We train the four prefixes of $(p, p_r, \mathcal{R})$ needed along this path, with the source image always present. The three dropout cases remove reference visuals alone, reference text together with visuals, or all three groups. Together with the full-condition case, these yield the four states in Eq. (4). Whenever an earlier group is dropped, all subsequent groups are dropped with it. Let $q_{\mathrm{tar}}, q_{\mathrm{txt}}, q_{\mathrm{vis}}$ be the hierarchical dropout probabilities. The probabilities of the four states are

$$\begin{aligned} \pi_0 &= q_{\mathrm{tar}}, & \pi_1 &= (1-q_{\mathrm{tar}})q_{\mathrm{txt}}, \\ \pi_2 &= (1-q_{\mathrm{tar}})(1-q_{\mathrm{txt}})q_{\mathrm{vis}}, & \pi_3 &= (1-q_{\mathrm{tar}})(1-q_{\mathrm{txt}})(1-q_{\mathrm{vis}}). \end{aligned} \tag{A.3}$$

Our Stage III configuration uses $(q_{\mathrm{tar}}, q_{\mathrm{txt}}, q_{\mathrm{vis}}) = (0.1, 0.167, 0.267)$, giving $(\pi_0, \pi_1, \pi_2, \pi_3) \approx (0.10, 0.15, 0.20, 0.55)$. One state is sampled for each training batch. This nested sampling exposes the model to the four predictions required by the guidance rule.

An absent text condition is obtained by encoding an empty string. For the reference visual group, zero-pixel inputs are passed through the VAE and image encoder to obtain null video, first-frame, and image-embedding conditions. The source image's latent conditions and image embedding are retained in every state. Thus c_0 still anchors generation to the source image.

Guidance interpretation. When $s_{\mathrm{tar}} = s_{\mathrm{txt}} = s_{\mathrm{vis}} = s$, Eq. (5) telescopes to $v_0 + s(v_3 - v_0)$, recovering joint CFG between the same endpoint states. Distinct scales allow each incremental contribution to receive a different weight. Each difference compares two predictions that share all preceding conditions and differ by one added group. The increments depend on the specified conditioning order and can contain overlapping semantic information. The four states can be evaluated in batches; their count alone does not determine wall-clock latency.

A.3 Distillation Objectives and Sampling

Score relation and student update. For $y_\sigma = (1-\sigma)x + \sigma\eta$ with $0 < \sigma < 1$, a clean estimate d corresponds to the score estimate

$$\widehat{s}(d) = \frac{(1-\sigma)d - y_\sigma}{\sigma^2}, \qquad \widehat{s}(d_{\mathrm{F}}) - \widehat{s}(d_{\mathrm{T}}) = \frac{1-\sigma}{\sigma^2}(d_{\mathrm{F}} - d_{\mathrm{T}}). \tag{A.4}$$

DMD estimates a reverse-KL distribution-matching gradient from this difference (Yin et al., 2024b;a). The normalized surrogate in Eq. (7) expresses this update through clean predictions. Its per-sample normalizer $\rho = D^{-1}\|x - d_{\mathrm{T}}\|_1$ averages the absolute difference over all target latent elements. For each sample, the gradient with respect to x is Δ/D. Only the generated prediction carries gradients to the student; the constructed target, including both score estimates and the normalizer, is detached. Normalization is evaluated in FP32, with `nan_to_num` applied to degenerate quotients.

Auxiliary score learning. The fake-score network minimizes Eq. (8) on detached student predictions, using fresh noise levels and perturbations to construct y_σ. The teacher remains frozen, while all DiT parameters of the student and auxiliary network are updated. We use these distribution-matching and fake-score objectives without an adversarial term.

Multi-step simulation and inference. Following DMD2 (Yin et al., 2024a), we simulate the student's sampling process to construct its training inputs. For $1 = t_1 > \cdots > t_K > t_{K+1} = 0$, start from $z_{t_1} \sim \mathcal{N}(0, \mathbf{I})$ and alternate

$$\widehat{z}_0^{(k)} = z_{t_k} - t_k v_\theta(z_{t_k}, t_k; c_3), \qquad z_{t_{k+1}} = (1-t_{k+1})\widehat{z}_0^{(k)} + t_{k+1}\epsilon_k, \tag{A.5}$$

where the ϵ_k are independent Gaussian draws. Training uniformly selects $k \in \{1, \ldots, K\}$, simulates earlier steps without gradients, and uses $x = \widehat{z}_0^{(k)}$ for the distribution-matching update. Inference executes all $K = 4$ steps and decodes the final clean prediction with the VAE.

Our student uses noise levels $(t_1, \ldots, t_5) \approx (1, 0.904762, 0.75, 0.513158, 0)$. The teacher guidance configuration is $(s_{\mathrm{tar}}, s_{\mathrm{txt}}, s_{\mathrm{vis}}) = (4, 5, 2)$; the student and fake-score network use full conditioning. The auxiliary network receives five updates per student update, with fresh generated samples.

Table 3: Judge dimensions. Higher scores are better in both rubrics, including the inverted reference-leakage score (RL). The two temporal dimensions share an abbreviation but use different prompts and scales.

Dimension	Evaluation focus
Direct VLM: separately assigned integer scores, 0–5	
Effect Fidelity (EF)	Correct effect, affected subject, strength, essential stages, temporal order, and final state.
Source Preservation (SP)	Source identity and non-target content, allowing effect-required changes.
Temporal Consistency (TC)	Smooth progression and completion; flicker, popping, tearing, or discontinuities.
Reference Disentanglement (RD)	Transfer of the effect without copying incidental reference appearance or scene content.
Visual Quality (VQ)	Clarity, geometry, structural integrity, boundaries, lighting, and physical plausibility.
Edit Locality (EL)	Changes confined to the intended content, with unrelated regions remaining stable.
Evaluation agent: continuous scores, 0–1	
Semantic Fidelity (SF)	Presence, strength, spatial coverage, and completion of the demonstrated effect.
Subject–Reference Disentanglement (SRD)	Source-grounded subject identity through the transformation, avoiding reference-subject cloning.
Temporal Coherence (TC)	Effect onset and progression, temporal artifacts, and subject continuity.
Reference Leakage (RL)	Absence of unintended reference-specific objects, textures, colors, or shapes.

For the student update, score-evaluation noise levels are sampled from the central 96% of the 1,000-level training schedule with shift 3; auxiliary updates use the full schedule. The deployed student uses the fixed guidance behavior learned during distillation.

B Evaluation Details

B.1 Test Cases and Scoring Dimensions

Fixed test inputs. The fixed test set contains 2,004 cases and 1,439 distinct reference videos, covering diverse subjects and scenes. For each case, a target video supplies the source image through its first frame, and a distinct video with the same effect label supplies the reference. The evaluation uses the source frame, reference video, and captions; the remaining frames of the target video are not provided to the judges. All methods receive the same source images, reference videos, and captions through their respective conditioning interfaces.

Training–test separation. For every evaluation case, both the target video supplying its source image and the reference video are excluded from all four stages of Reveris training. They serve neither as supervised training targets nor as reference conditions in Stages I–III, and are also excluded from the conditioning examples used for DMD in Stage IV. This separation ensures that evaluation measures transfer to unseen video content.

What each dimension measures. Table 3 defines the two judge rubrics. Preservation is assessed relative to the intended effect: a transformation may legitimately change a subject's shape, material, or category. For such cases, the rubric checks the initial source identity and whether subsequent changes follow the effect, while treating unrelated subject or scene copying as leakage. Dimensions retain separate scores, so strong preservation or visual quality cannot conceal failure to realize the effect.

B.2 Direct VLM Assessment

Primary evaluator and inputs. The main direct-VLM results use GPT-5.6 Sol. The evaluator receives the source image, sampled reference and generated frames, and task descriptions, with each visual input labeled by its role. Both videos are sampled at 2 fps with the last frame included; timestamps and durations preserve their temporal context. Frames are supplied at native resolution as quality-95 JPEGs with high image detail. We use high reasoning effort, an 8,192-token response limit, and default temperature. Prompts, input preparation, and inference settings are fixed across generation methods.

Scoring rubric. The prompt first asks the VLM to identify the transferable effect, incidental reference content, source invariants, and observed output progression, then score all six dimensions with visible evidence. The shared anchors are 5 for success with at most minor defects, 4 for overall success with non-critical defects, 3 for partial success with clear deficiencies, 2 for weak realization, 1 for severe failure, and 0 for an absent, opposite, copied, or unusable result. EF is capped at 1 when the effect is absent, fundamentally incorrect, or applied to the wrong subject.

Alternative evaluators. We also score the same generated outputs with Gemini 3 Flash Preview and Seed 2.0 Lite (version 260428). Gemini uses the same frame inputs, prompt, and response limit as GPT-5.6 Sol, with maximum reasoning effort and default temperature. Seed uses the same six-axis rubric with native video input at 2 fps and temperature 0. The comparison therefore measures robustness across evaluator configurations, including their input formats and inference settings.

B.3 Evidence-Grounded Evaluation Agent

Evaluating visual effect transfer requires judging both the intended transformation and its grounding in the source content. A nearly static output may preserve the source yet fail to realize the effect; a successful material or category transformation may substantially change its appearance. Our agent makes the evidence behind these judgments explicit through three stages: *blind observation*, *grounded assessment*, and *targeted review* (Figure 7). This organization follows recent work on staged video evaluation and tool-assisted assessment (Han et al., 2025; Zhang et al., 2025a). Seed 2.0 Lite carries out all three stages, with six automatic evidence tools and optional visual inspection actions.

B.3.1 Blind Observation: Establishing What Happened

The first stage receives the generated video Y and four output-only tool reports: feature drift, perceptual change, optical flow, and subject tracking. The source image I, reference video R, target caption p, and effect metadata are withheld. The judge describes the initial subject, visible changes, estimated onset, completion, and observable anomalies. If a particular interval is unclear, the judge can request a temporal replay and update its observation.

These observations anchor the subsequent assessment in what the output actually depicts. For example, an observation of an unchanged subject remains evidence against effect realization even when the target caption describes a dramatic transformation. The first stage assigns no quality scores; task-dependent interpretation begins in the assessment stage.

B.3.2 Grounded Assessment: Relating Changes to the Task

The second stage receives Y, R, the effect metadata (name, description, and onset clause when available), the target caption, the initial observations, and all six tool reports. Two additional tools supply evidence about content origin. *Subject similarity* compares generated subject regions with the source subject and the reference's initial subject, using normalized DINOv2 features. It reports both cosine-similarity curves and their source-minus-reference difference. Source-image evidence reaches the VLM through this tool. *The leakage probe* searches the output for reference-specific entities, returning their detection scores, sampled-frame locations, and bounding boxes. Together, these tools help distinguish retaining the source subject from introducing incidental reference content: the correct source subject can still carry an unwanted reference object or texture.

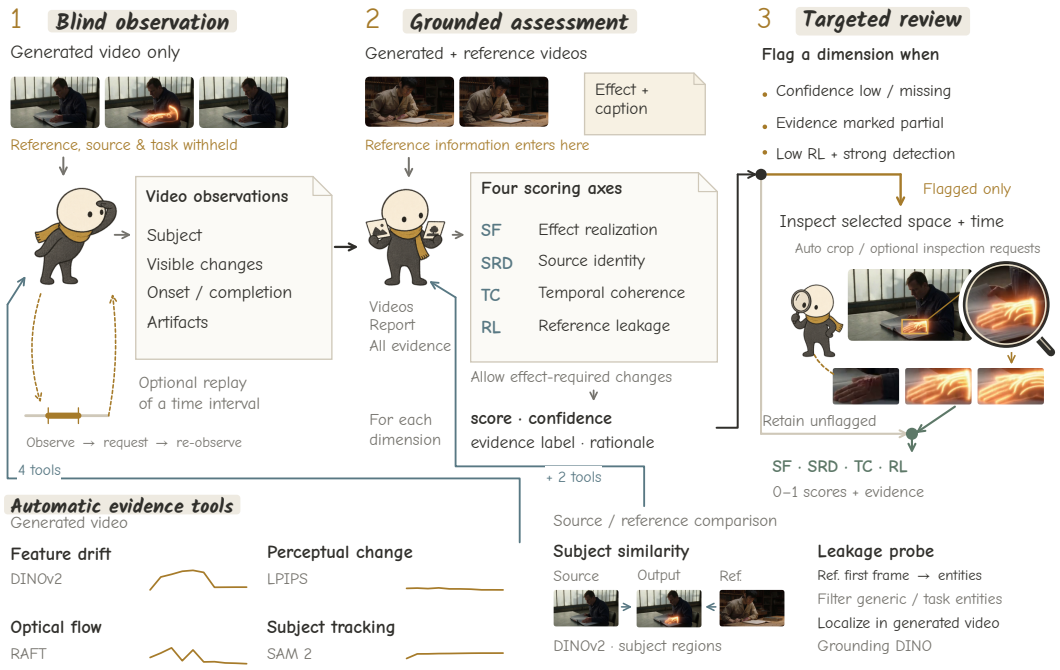

Figure 7: Evidence-grounded evaluation agent. The same VLM first observes the generated video with task and reference information withheld, then assesses transfer using the reference and tool evidence, and revisits flagged dimensions through localized inspection. Unflagged assessments remain unchanged. The lower branch separates output-only tools from source/reference comparison tools; source-image evidence enters through subject similarity. Higher RL indicates less leakage.

The rubric interprets these cues relative to the effect. For a category-changing transformation, it checks source identity at the start and the intended transition thereafter; declining appearance similarity alone does not establish reference-subject copying. SF concerns effect presence, strength, coverage, and completion; SRD concerns source-grounded subject identity; TC concerns progression and temporal stability; RL concerns incidental reference contamination. For each dimension, the judge provides a score, confidence estimate, evidence label, and rationale. Scores lie in $[0, 1]$, with 1, 0.5, and 0 anchoring successful, partial, and failed realization of the criterion. The evidence label is one of `support`, `partial`, `contradict`, or `missing`; each rationale refers to an observation or tool measurement. Confidence is the judge's self-reported certainty and is used to route review, without weighting the reported scores.

B.3.3 Targeted Review: Resolving Uncertain Evidence

Review selection. A dimension is flagged when its assessment or confidence is missing, its confidence is below 0.7, or its evidence label is `partial`. RL is additionally flagged when its initial score is below 0.5 and the leakage probe supplies a bounding box with detection confidence at least 0.6, even if the judge declares high confidence. This additional check addresses a specific failure mode: a strong object-detection response can lead to a leakage penalty for an entity that is absent from the video. If no dimension is flagged, the initial assessments are final.

Leakage evidence and visual verification. The initial scoring prompt uses three detection bands. A peak below 0.4 suggests RL of at least 0.85; a peak in $[0.4, 0.6)$ is treated as ambiguous, with RL below 0.7 requiring visible confirmation; a peak of at least 0.6 suggests RL at most 0.4 and leads to the additional review rule above. These bounds are instructions to the initial judge. During review, a flagged RL dimension receives an automatic close-up when a localized detection of confidence at least 0.4 is available. We crop the strongest detection's region over a window extending 0.6 s before and after its sampled time, clipped to the video boundaries. The reviewer checks whether the named entity is visible in this clip and the full video, and may revise the initial penalty accordingly.

Table 4: Automatic evidence tools. The first four inspect only the generated video and precede blind observation; the last two add source/reference evidence before assessment. Rates and frame caps refer to tool inputs, separately from VLM sampling.

Tool	Evidence and implementation	fps / cap
Feature drift	DINOv2-base (Oquab et al., 2024): distance of normalized frame features from the generated first frame.	4 / 32
Perceptual change	LPIPS-AlexNet (Zhang et al., 2018): adjacent-frame differences and peaks above 0.3.	8 / 32
Optical flow	RAFT-Small (Teed & Deng, 2020): motion magnitude and its temporal variation.	4 / 32
Subject masks	SAM 2 Hiera Tiny (Ravi et al., 2025): subject area and centroid over sampled frames.	4 / 32
Subject similarity	DINOv2-base cosine similarities between generated subject crops and source/reference first-frame crops.	4 / 32
Leakage probe	Grounding DINO Tiny (Liu et al., 2024): localized detections of reference-specific entities.	2 / 8

Inspection actions and stopping. Recent visual reasoning systems make spatial and temporal inspection explicit (Su et al., 2025; Zheng et al., 2026b). The agent can replay a temporal interval during observation and inspect a spatially cropped interval during review. It examines the selected clips alongside the original stage inputs. Observation and review each allow at most two additional calls, with at most two requested clips executed per round; an empty action list ends inspection early. Automatic leakage verification is attached before the first review call and does not consume this optional action budget.

Selective score revision. Review re-examines only the flagged dimensions using their initial assessments and the additional visual evidence. The prompt asks it to retain the score when evidence remains ambiguous and explain that ambiguity. Let q_d^{cmp} and q_d^{rev} denote the initial and reviewed assessments for dimension d, and let $\mathcal{U}$ contain the flagged dimensions. The final assessment is

$$q_d^{\mathrm{final}} = \begin{cases} q_d^{\mathrm{rev}}, & d \in \mathcal{U} \text{ and a reviewed assessment is available,} \\ q_d^{\mathrm{cmp}}, & \text{otherwise.} \end{cases} \tag{B.1}$$

Assessments of unflagged dimensions remain unchanged. The four final scores are reported separately.

Review example. The inspection in Figure 7 illustrates this process on a generated clip in which a luminous contour appears on a man's hand and forearm, then fades. The reference-entity detector proposes a *chisel* with confidence 0.643, and the initial judge assigns RL $= 0.300$. The resulting review inspects a crop of the detected region over 0–1.1 s, finds no visible chisel, and revises RL to 1.000. SF, SRD, and TC retain their initial assessments. The sequence illustrates how localized visual evidence can overturn a detector-induced penalty.

B.3.4 Tool Configuration and Reproducibility

Evidence extraction. Table 4 lists the six tools and their sampling budgets. Feature drift is the Euclidean distance between normalized DINOv2 frame features and the generated first-frame feature. The tool summarizes its final 20% as a plateau estimate and reports the first crossing of 30% of that estimate when the plateau exceeds 10^{-3}. Subject masks start from a center-point prompt and use the previous mask's padded bounding box to prompt the next sampled frame, returning to the center prompt if the mask is lost. Similarity features use subject boxes padded by 10%, with a full-frame fallback when a usable box is unavailable. For leakage detection, a separate call to the same VLM extracts up to eight candidate entities from the reference first frame. Entities appearing in the target caption and generic terms for body parts, clothing, scenery, lighting, and broad subject classes are filtered before detection. The judge receives curve summaries and detection results as structured text; these measurements provide cues whose interpretation also depends on the visible effect and motion.

Inference settings. All agent stages use 4 fps VLM video sampling and temperature 0.3. Response budgets are 900 tokens for observation, 1,500 for assessment or review, and 200 for reference-entity extraction. Inspection clips are sampled at 8 fps with at most 32 frames; crops are enlarged twofold with bicubic interpolation before the usual VLM video sampling.

B.4 VBench, Aggregation, and Inference Timing

VBench on the fixed test inputs. We compute the five quality dimensions in Section 5 on the generated videos using the official VBench implementations (Huang et al., 2024). Each metric retains its official decoding and preprocessing; imaging quality uses the default setting that limits the longer side to 512 pixels. Subject and background consistency retain the official frame-pair-weighted aggregation, while motion smoothness, aesthetic quality, and imaging quality use video means. Per-video MUSIQ values are divided by 100 to match VBench's aggregate imaging-quality scale. These are VBench component metrics on our transfer inputs, distinct from the standard VBench prompt-suite leaderboard.

Reporting. The main results cover all 2,004 cases: direct and agent scores use an equal-weight case mean on their original scales, and VBench follows the aggregation rules above. Cross-evaluator comparisons use the cases with successful scores from every evaluator. All dimensions are reported separately.

Average rank. Within each direct evaluator, we rank the same eight methods in Table 1 by their unrounded case-mean score on each of EF, SP, TC, RD, VQ, and EL. Rank 1 is best, and exact ties receive the average of the occupied ranks. Avg. Rank is the equal-weight mean of these six ranks, computed separately for each evaluator. It summarizes relative standing among the listed methods; the accompanying dimension scores retain the size of performance differences.

Inference latency. We measure ten fixed cases three times each after two warm-up runs per model on each active GPU worker. Each video contains 49 frames at 832×480 and 16 fps and is generated on one NVIDIA A100 80GB GPU with CPU offloading. CUDA synchronization brackets resizing, text/image/reference encoding, denoising, VAE decoding, and conversion to CPU uint8 frames. Model loading, input reading and validation, and MP4 writing are excluded. For each repetition, we reset the sampling seed and recompute the conditions. Methods retain native precision, samplers, CFG, VAE tiling, and offloading implementations, with TF32 disabled. We use eight independent GPUs on a shared host, each generating one video at a time. We average repetitions within cases, then across cases. Conditioned denoiser evaluations count each guidance branch, including those batched into one forward call, separately from sampling steps and forward calls.

C Additional Experimental Results

C.1 Implementation and Sampling Configurations

Four-stage training configuration. All four stages use a per-device micro-batch of one target–reference pair. Training clips contain 49 frames at 832×480, and we use BF16 mixed precision. Stages I, III, and IV use an initial learning rate of 10^{-6}, including both the student and auxiliary score network in Stage IV. Stage II uses 10^{-4} for the gate parameters and 10^{-3} for the signed amplitudes. Table 5 summarizes the training updates and parallel configuration; gradient accumulation is one throughout. Stages I–III each use 64 A100 GPUs. Stage IV uses 8-way sequence parallelism within each of ten data-parallel replicas, giving a global batch of ten across 80 A100 GPUs.

Compared implementations. VAP uses the released Wan2.1-based model with guidance 5. VFX-Master uses its released CogVideoX-based in-context conditioning checkpoint after 40,000 training updates, with dynamic text guidance 6 and no per-case adaptation. RefVFX uses the public reimplementation linked by its project, with Wan2.1-FLF2V-14B-720P and the released step-10,000 adapter.[2] We evaluate it with 50 sampling steps, text guidance 6, and reference guidance 2. Re-

[2]The implementation identifies itself as an unofficial reimplementation: `https://github.com/maxwelljones14/refVFX`.

Table 5: Four-stage training settings. DP and SP denote data and sequence parallelism. Updates in Stage IV count student updates; the auxiliary score network receives five updates per student update.

Stage	Updates	A100 GPUs	DP	SP	Global batch
I: Reference adaptation	6,000	64	64	1	64
II: Modulation calibration	6,000	64	64	1	64
III: Multi-source guidance	6,000	64	64	1	64
IV: Few-step distillation	600	80	10	8	10

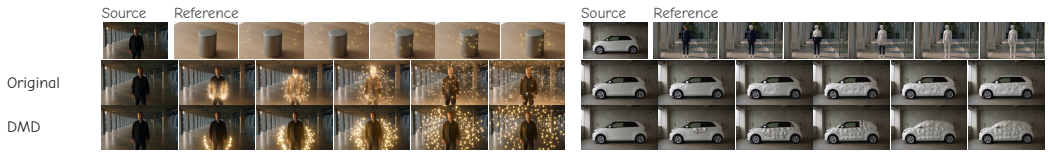

Figure 8: The 50-step teacher (Original) and four-step student (DMD) on firefly swarm and origami faceting. Each pair uses matching inputs and frame indices.

fVFX+CausVid adds the CausVid adapter and uses six steps with the same guidance scales. The RefVFX interface conditions on the source image, reference video, and target text; it does not separately consume the reference caption.

Reveris and the conditioning variants. The teacher uses 50 steps with the three scales $(4, 5, 2)$; the student uses four fully conditioned steps of Eq. (A.5). The modulation variants in Table 1 are separately trained models with the specified conditioning designs. The model without the prior retains the designated reference-interaction blocks, while the model without gating learns signed amplitudes for the temporal-variation prior.

Distillation configuration. The student and auxiliary score network are initialized from the Stage III teacher. We use weight decay 10^{-4}, cosine learning-rate scheduling with 100 warm-up steps, and five auxiliary updates per student update. The results in this paper use the checkpoint after 600 student updates. All DiT parameters of both networks are trainable; the teacher remains frozen. The student and auxiliary network optimize Eqs. (7) and (8), respectively. For *Reveris w/o prior + DMD*, we apply the same distillation procedure to the interaction-only model, initializing its student and auxiliary network from that teacher. Both distilled models are evaluated after 600 student updates with four fully conditioned sampling steps.

Inference cost. In Table 2, each CFG branch counts as a conditioned evaluation, including VFX-Master's two branches batched into one forward call. The ratio of mean teacher latency to mean student latency is 22.815, with a paired 95% bootstrap interval of $[22.680, 22.948]$ over the ten case means. Each case mean averages three measured repetitions.

C.2 Paired Distillation Analysis

Using GPT-5.6 Sol scores, we group the 2,004 matched teacher/student outputs by their 1,439 distinct reference videos. Each bootstrap replicate samples 1,439 reference-video groups with replacement, retaining all cases and paired method scores within each selected group. We then compute the student-minus-teacher mean difference over the resampled cases, giving every case equal weight. Table 6 reports the 2.5th and 97.5th percentiles over 5,000 replicates under the fixed judging protocol. The intervals quantify variation across reference-video groups, conditional on the generated outputs and evaluator scores. Figure 8 complements these aggregate scores with matched generations: both models produce an emerging firefly swarm around the person and progressive faceting of the car's surface.

Reference conditioning after distillation. We apply the same DMD procedure to the model without the prior and compare its four-step student with Reveris after 600 student updates. Table 7 reports all six GPT-5.6 Sol dimensions on 2,004 cases, with the same paired reference-group bootstrap used

Table 6: Paired student-minus-teacher GPT-5.6 Sol differences on all 2,004 cases with 95% percentile bootstrap intervals. We resample reference-video groups with replacement 5,000 times, retaining all paired cases within each selected group.

Metric	Teacher	Student	Difference [95% CI]
EF	3.073	3.110	+0.037 [-0.007, +0.082]
SP	3.628	3.700	+0.072 [+0.032, +0.113]
VQ	3.249	3.392	+0.143 [+0.114, +0.174]

Table 7: Reference conditioning after DMD on 2,004 cases, evaluated by GPT-5.6 Sol (0–5). Both models use four sampling steps after 600 student updates. Differences are Reveris minus w/o prior + DMD; brackets give paired 95% intervals from 5,000 reference-group bootstrap replicates.

Metric	Reveris w/o prior + DMD	Reveris (Ours)	Difference [95% interval]
EF	2.935	**3.110**	+0.176 [+0.136, +0.215]
SP	3.698	**3.700**	+0.001 [-0.034, +0.037]
TC	3.365	**3.468**	+0.103 [+0.068, +0.138]
RD	**4.988**	4.982	-0.005 [-0.013, +0.002]
VQ	3.287	**3.392**	+0.104 [+0.075, +0.132]
EL	3.878	**3.892**	+0.014 [-0.032, +0.058]

above. Adaptive conditioning raises EF from 2.935 to 3.110, TC from 3.365 to 3.468, and VQ from 3.287 to 3.392; the paired 95% intervals are positive for all three gains. SP, RD, and EL differ by at most 0.014, with intervals that include zero. The EF advantage is 0.146 before distillation and 0.176 afterward, showing that stronger reference fidelity persists when both conditioning designs receive the same DMD procedure.

C.3 Robustness to Evaluator Choice

Our main comparison uses GPT-5.6 Sol as the direct evaluator. Following recent evaluations that check conclusions with alternative judges (Qian et al., 2025; Lei et al., 2026), we examine whether the method rankings and teacher–student trends persist under Gemini 3 Flash Preview and Seed 2.0 Lite. All three apply the six-axis rubric to the same generated outputs; their configurations are specified in Section B.2.

Common evaluation set. GPT-5.6 Sol and Seed score all 2,004 cases for each of the eight methods in Table 1. Gemini has valid scores for 2,002 cases per method, with the same two cases missing throughout. All comparisons in this subsection use the 2,002 cases scored by all three evaluators for every method. Table 8 reports the alternative-evaluator means on this common set.

Stable comparisons across different score scales. Among these eight methods, Reveris ranks first on all six dimensions under every evaluator, giving Avg. Rank 1.00 for GPT-5.6 Sol, Gemini, and Seed 2.0 Lite. The EF rankings agree exactly: model-level Spearman ρ is 1.000 for every evaluator pair, with reference-group bootstrap 95% intervals of $[0.952, 1.000]$. SP, TC, and VQ correlations are at least 0.952; EL ranges from 0.857 to 0.952 (Table 9). The adaptive teacher exceeds both modulation ablations on EF, SP, and VQ under every evaluator. Thus, the comparisons support adaptive reference conditioning and the four-step model across evaluator choices despite differences in absolute score levels.

Method rankings and individual scores. Model-level correlation compares eight method means, whereas video-level correlation compares 16,016 paired scores before aggregation. We report both, together with quadratic-weighted Cohen's κ for agreement on the ordinal 0–5 scale. Video-level EF correlations range from 0.583 to 0.720: stable method rankings do not imply interchangeable individual judgments. These analyses establish cross-evaluator consistency, rather than alignment with human ratings.

Table 8: Alternative Direct VLM scores on the same 2,002 cases for all eight methods. Higher 0–5 scores and lower Avg. Rank are better; best and second-best displayed values within each evaluator are bold and underlined. Avg. Rank averages the six dimension ranks.

Method	EF	SP	TC	RD	VQ	EL	Avg. Rank↓
Gemini 3 Flash Preview							
VAP	3.107	3.531	3.178	4.898	3.211	3.639	7.50
VFXMaster	3.778	3.734	3.321	4.885	3.170	4.038	5.83
RefVFX	3.529	3.665	3.277	4.932	3.165	3.870	6.83
RefVFX+CausVid	3.683	3.960	3.593	4.938	3.396	4.046	4.00
Reveris w/o prior	3.764	3.935	3.623	4.935	3.494	4.133	3.67
Reveris w/o gate	3.230	3.804	3.429	<u>4.948</u>	3.328	3.980	5.00
Reveris Teacher	<u>3.893</u>	<u>4.098</u>	<u>3.737</u>	4.942	<u>3.662</u>	<u>4.217</u>	<u>2.17</u>
Reveris (Ours)	**3.955**	**4.221**	**3.874**	**4.960**	**3.843**	**4.360**	**1.00**
Seed 2.0 Lite							
VAP	2.813	3.889	4.019	<u>4.998</u>	3.832	3.832	6.83
VFXMaster	3.404	4.158	4.168	4.993	3.781	4.269	5.67
RefVFX	2.972	4.041	4.135	<u>4.998</u>	3.739	4.092	6.17
RefVFX+CausVid	3.110	4.211	4.347	4.997	3.868	4.098	4.67
Reveris w/o prior	3.357	4.221	4.385	4.996	3.995	4.319	3.50
Reveris w/o gate	2.965	4.141	4.310	4.995	3.940	4.287	5.33
Reveris Teacher	<u>3.527</u>	<u>4.338</u>	<u>4.488</u>	4.994	<u>4.114</u>	<u>4.411</u>	<u>2.83</u>
Reveris (Ours)	**3.700**	**4.435**	**4.614**	**4.999**	**4.269**	**4.556**	**1.00**

Table 9: Cross-evaluator agreement on the common 2,002 cases. Method-level Spearman ρ uses eight method means; video-level ρ and quadratic-weighted Cohen's κ use all 16,016 paired output scores. Sol, Gemini, and Seed denote GPT-5.6 Sol, Gemini 3 Flash Preview, and Seed 2.0 Lite.

Evaluator pair	EF	SP	TC	RD	VQ	EL
Method-level Spearman ρ						
Sol / Gemini	1.000	0.952	0.976	0.690	0.976	0.857
Sol / Seed	1.000	1.000	0.976	0.405	0.952	0.952
Gemini / Seed	1.000	0.952	1.000	0.238	0.976	0.905
Video-level Spearman ρ						
Sol / Gemini	0.602	0.553	0.497	0.187	0.533	0.526
Sol / Seed	0.720	0.611	0.484	0.149	0.550	0.592
Gemini / Seed	0.583	0.541	0.461	0.125	0.528	0.501
Video-level weighted κ						
Sol / Gemini	0.498	0.532	0.471	0.204	0.476	0.547
Sol / Seed	0.672	0.523	0.284	0.148	0.369	0.553
Gemini / Seed	0.583	0.558	0.363	0.089	0.465	0.578

Reference disentanglement and score saturation. RD receives the maximum score on 97.98% of outputs under GPT-5.6 Sol, 95.61% under Gemini, and 99.79% under Seed. Its model-level correlations are consequently less stable (0.238–0.690), and individual-score correlations are low (0.125–0.187). RD retains equal weight in Avg. Rank; our interpretation of transfer quality centers on the non-saturated EF, SP, and VQ dimensions.

Distillation under alternative evaluators. Table 10 compares the student and teacher on the common cases, using 5,000 paired bootstrap replicates over their 1,438 reference-video groups. SP and VQ improve under all three evaluators, with positive 95% intervals. Mean EF is also higher under each evaluator, although its intervals include zero for GPT-5.6 Sol and Gemini. Together with the $22.8\times$ latency reduction, these results support few-step transfer that retains strong effect fidelity while improving source preservation and visual quality. Bootstrap resampling preserves paired method and evaluator scores within each reference group, with generated outputs and judgments held fixed.

Table 10: Student-minus-teacher differences across evaluators on the same 2,002 cases. Brackets give paired 95% intervals from 5,000 reference-video-group bootstrap replicates.

Evaluator	EF	SP	VQ
GPT-5.6 Sol	+0.038 [-0.007, +0.083]	+0.072 [+0.033, +0.114]	+0.143 [+0.114, +0.173]
Gemini	+0.062 [-0.008, +0.131]	+0.123 [+0.065, +0.183]	+0.181 [+0.128, +0.234]
Seed 2.0 Lite	+0.172 [+0.114, +0.232]	+0.097 [+0.053, +0.142]	+0.155 [+0.117, +0.193]

C.4 Reference Modulation Diagnostics

We compare a separately trained mean-prior model with inference-time changes to the Stage III teacher. *Mean prior + learned gate* starts from the shared Stage I checkpoint and uses the mean reference-change prior throughout Stages II and III, retaining the learned signed amplitudes and target-conditioned gates. The mean is computed separately for each reference video and broadcast over its tokens. For the inference-time controls, *Mean prior* replaces the trained teacher's normalized reference-change map with its average over reference tokens. *Shuffled prior* permutes the token positions while preserving the map's values; we show three permutations with seeds 101, 202, and 303. The inference-time controls keep all teacher parameters fixed. All rows use the same source image, target prompt, paired reference video and caption, sampling seed, 50 sampling steps, and guidance scales $(4, 5, 2)$.

Figure 9 shows three examples. During partial mechanization, the original map better preserves the dog's floppy ears and head shape than the inference-time mean and shuffled priors, which produce pointed ears and a more angular robotic head. For object orbit, these inference-time controls introduce elongated dark bands; the shuffled variants also deform the cups. In clone solidification, the original produces a more solid frame toward the end, while these controls retain luminous outlines. These differences recur across the three permutations, illustrating sensitivity to the spatial and temporal placement of reference-change cues in these examples. The separately trained mean-prior model also introduces additional orbiting objects and retains bright outlines in the clone-solidification example.

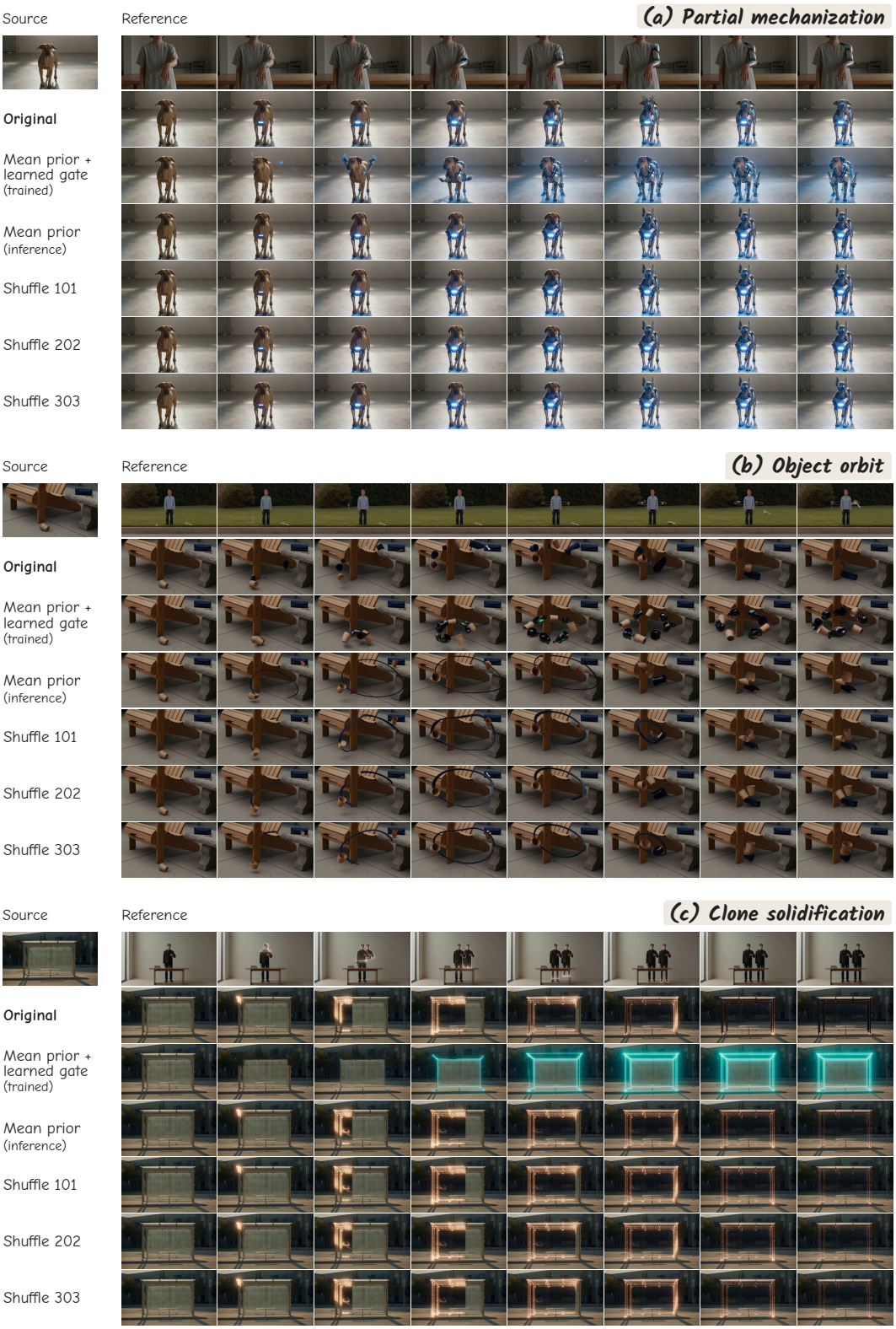

Figure 9: Reference-modulation comparisons. Mean prior + learned gate is trained through Stages II–III from shared Stage I; Mean prior and Shuffle modify the teacher at inference. Rows show eight matched frames.

C.5 Paired Reference Control

Figure 5 compares two generations from the same source image and target prompt. Both use the four-step student with sampling seed 42 and identical inference settings, producing 49 frames at 832×480 and 16 fps. The changed condition is the reference video together with its corresponding caption: one depicts leftward flight over a field, and the other depicts rightward flight from a snowy clearing. This preserves the model's paired video–caption input interface. The target description specifies the subject, setting, takeoff, and horizontal flight while leaving the left/right direction open. The generated woman transitions from standing to opposite head-first flight directions while retaining her source clothing and outdoor setting. Thus, this example illustrates how the paired reference conditions steer the realization of a shared target description.

The figure shows output frames 0, 12, 24, 36, and 48 for both generations, and reference frames sampled uniformly over each full clip. The exact target and reference descriptions are reproduced below.

Shared target prompt. A woman with light brown hair tied back, wearing a light blue shirt, dark cropped trousers, and dark shoes, stands on a garden lawn. She lifts off the ground, her body gradually tilting and extending from an upright stance into a nearly horizontal, head-first flying posture with her legs trailing behind. She flies through the garden with clear sideways displacement, remaining visibly separated from the grass below. Her shirt and hair move gently with her flight. Trees, a low brick wall, and a house shift relative to her as she moves through the outdoor space. The takeoff and sustained airborne movement unfold in one continuous shot.

Reference caption for leftward flight. A young woman wearing a long white sleeveless wedding dress, a small tiara, and white shoes stands among tall golden grasses at sunset. She holds a bouquet with a long red ribbon. The woman leans toward the left side of the image and reaches forward as her feet lift clear of the grass. Her body extends into a nearly horizontal flying posture, with her head leading toward screen left and her feet trailing on the right. She glides leftward above the field, her dress and ribbon streaming behind her. The camera tracks her movement while the low sun and open grassland remain visible.

Reference caption for rightward flight. A young woman with short dark hair and a dark cap stands in a snowy outdoor clearing, wearing a white top beneath an open camouflage jacket. She leans forward toward the right side of the image and extends her arms slightly behind her. Her body lifts off the ground and straightens into a horizontal flying posture. She flies head-first toward screen right, with her legs trailing toward the left, and gains height as the camera tracks alongside her. Her jacket and hair move with the flight. The snowy ground gives way to a view across evergreen treetops and distant hills as she continues into the air.

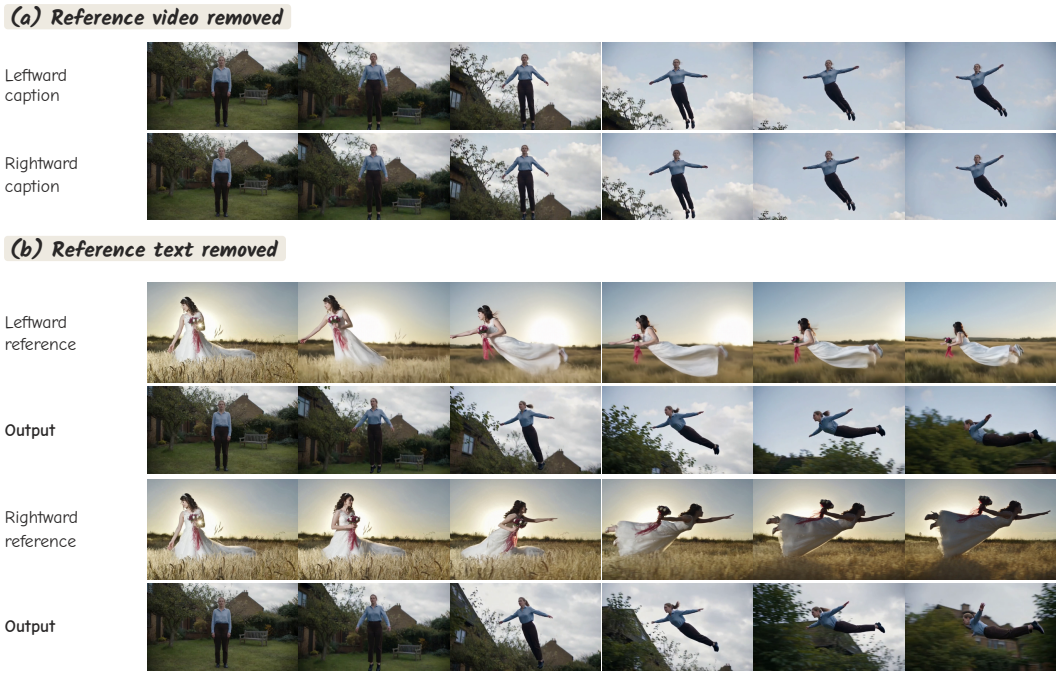

Figure 10: Two reference-input removal probes of the four-step student. (a) Left/right reference captions with reference visuals removed. (b) Opposite-flight reference videos with reference text removed. The probes share the source image, target prompt, and seed from Figure 5; their reference pairs are specified below.

C.6 Reference-Condition Removal

We examine whether a directional contrast remains when one part of the paired reference condition is absent. Both probes use the same four-step student, source image, target prompt, and seed as Section C.5, with output frames 0, 10, 19, 29, 38, and 48 shown in Figure 10. They are inference-time sensitivity tests of the fully conditioned student.

Removing reference visuals. We retain the two directional reference captions used in Figure 5 and replace the reference video's VAE latents, first-frame VAE condition, and image-encoder features with null conditions obtained by encoding zero-valued normalized pixel tensors. The reference token layout and textual conditioning remain intact. Despite the opposing directions specified in the captions, both outputs show left-leaning takeoff without the contrasting horizontal flight of the fully conditioned pair. Thus, the directional captions alone do not reproduce the paired-reference control observed in this example.

Removing reference text. This probe uses two flight references initialized from the same image: the original leftward-flying woman in a white dress and a generated rightward-flight counterpart. We encode an empty reference caption while retaining all reference visual conditions. Both outputs develop leftward horizontal flight despite the opposite reference directions. The pair differs from the field/snow references of Figure 5; it also yields leftward outputs with a shared, direction-neutral reference caption. The observation therefore concerns the absence of directional control from this reference pair without individually matched descriptions.

Together, these sequence-level observations support using the video and its matched caption as a joint control input. The student was distilled with full conditioning, and visual-only reference conditioning is outside the teacher's nested states; these probes characterize sensitivity at inference.

Table 11: Guidance-scale sensitivity of the 50-step teacher. Scales are ordered as $(s_{\text{tar}}, s_{\text{txt}}, s_{\text{vis}})$. Higher is better; best and second-best displayed values within each scoring panel are bold and underlined.

Guidance scales	EF	SP	TC	RD	VQ	EL
GPT-5.6 Sol (0–5), $n = 2{,}004$						
$(1, 1, 1)$	2.449	3.574	3.043	4.966	2.931	3.568
$(4, 4, 2)$	3.045	**3.632**	3.389	**4.985**	<u>3.227</u>	**3.744**
$(5, 5, 2)$	3.027	3.606	<u>3.397</u>	<u>4.977</u>	3.212	3.719
$(5, 5, 5)$	**3.099**	3.541	3.327	4.967	3.162	3.643
$(4, 5, 2)$ (default)	<u>3.073</u>	<u>3.628</u>	**3.407**	<u>4.977</u>	**3.249**	<u>3.734</u>

	Agent (0–1), $n = 2{,}004$				VBench (0–1), $n = 2{,}004$				
Guidance scales	SF	SRD	TC	RL	SC	BC	MS	AQ	IQ
$(1, 1, 1)$	0.647	<u>0.944</u>	**0.865**	**0.936**	**0.892**	**0.924**	**0.990**	0.549	0.621
$(4, 4, 2)$	0.816	**0.947**	<u>0.863</u>	<u>0.934</u>	<u>0.850</u>	<u>0.904</u>	0.986	0.558	<u>0.627</u>
$(5, 5, 2)$	<u>0.818</u>	**0.947**	0.861	<u>0.934</u>	0.844	0.901	0.986	**0.561**	**0.628**
$(5, 5, 5)$	**0.822**	0.941	0.837	0.933	0.839	0.898	0.982	0.556	0.626
$(4, 5, 2)$ (default)	0.811	**0.947**	<u>0.863</u>	**0.936**	<u>0.850</u>	<u>0.904</u>	<u>0.987</u>	<u>0.559</u>	<u>0.627</u>

C.7 Multi-Source Guidance Scales

Table 11 compares five guidance settings with fixed Stage III weights, 50 sampling steps, and matched evaluation inputs. We use GPT-5.6 Sol for VLM scoring, consistent with the main evaluation, alongside the agent and VBench; all panels cover 2,004 cases per setting. By Eq. (5), $(1, 1, 1)$ retains full conditioning without CFG extrapolation, whereas $(5, 5, 5)$ applies joint guidance between the endpoint states.

Guidance improves effect realization: GPT-5.6 Sol EF rises from 2.449 at $(1, 1, 1)$ to 3.073 at $(4, 5, 2)$. Increasing only reference-visual guidance, from $(5, 5, 2)$ to $(5, 5, 5)$, further raises EF from 3.027 to 3.099, but lowers SP, TC, and VQ; agent TC also falls from 0.861 to 0.837. Stronger reference-visual guidance therefore introduces a tradeoff between effect fidelity and preservation. The neighboring $(4, 4, 2)$ and $(4, 5, 2)$ are much closer: EF is 3.045 versus 3.073, TC is 3.389 versus 3.407, and VQ is 3.227 versus 3.249, while SP slightly favors $(4, 4, 2)$ (3.632 versus 3.628). Although $(4, 5, 2)$ has the highest mean GPT-5.6 Sol TC and VQ, the small gaps indicate similar aggregate performance in this neighborhood. Other metrics also favor different settings: $(5, 5, 5)$ leads agent SF, $(5, 5, 2)$ leads VBench AQ and IQ, and $(1, 1, 1)$ leads SC, BC, and MS despite weaker effect fidelity.

The selected sequences in Figure 11 show how guidance affects the spatial extent and completion of an effect, complementing the aggregate scores. For spike growth, $(4, 5, 2)$ covers the chair with spikes while preserving the surrounding floor; $(4, 4, 2)$ and $(5, 5, 2)$ spread spikes across the floor, and $(5, 5, 5)$ introduces repeating background patterns. For split and merge, $(4, 5, 2)$ produces two copies of the cat and reunites them, whereas $(4, 4, 2)$ and $(5, 5, 2)$ leave two cats at the end. For particle disappearance, the same two alternatives generate conspicuous particle clouds but retain the kettle; $(4, 5, 2)$ disperses it into particles and leaves an empty countertop. These examples illustrate differences in completing the demonstrated process and preserving unaffected content that are less apparent from the close mean scores. Together with the quantitative tradeoffs, they support retaining $(4, 5, 2)$ as our balanced teacher configuration for distillation.

C.8 Additional Qualitative Comparisons

Figure 11 presents the guidance comparisons discussed above, with matched inputs, sampling seeds, and output frame indices. Figures 12 and 13 provide six further baseline comparisons spanning material transformations, scene transitions, structured light, particles, and fluid motion.

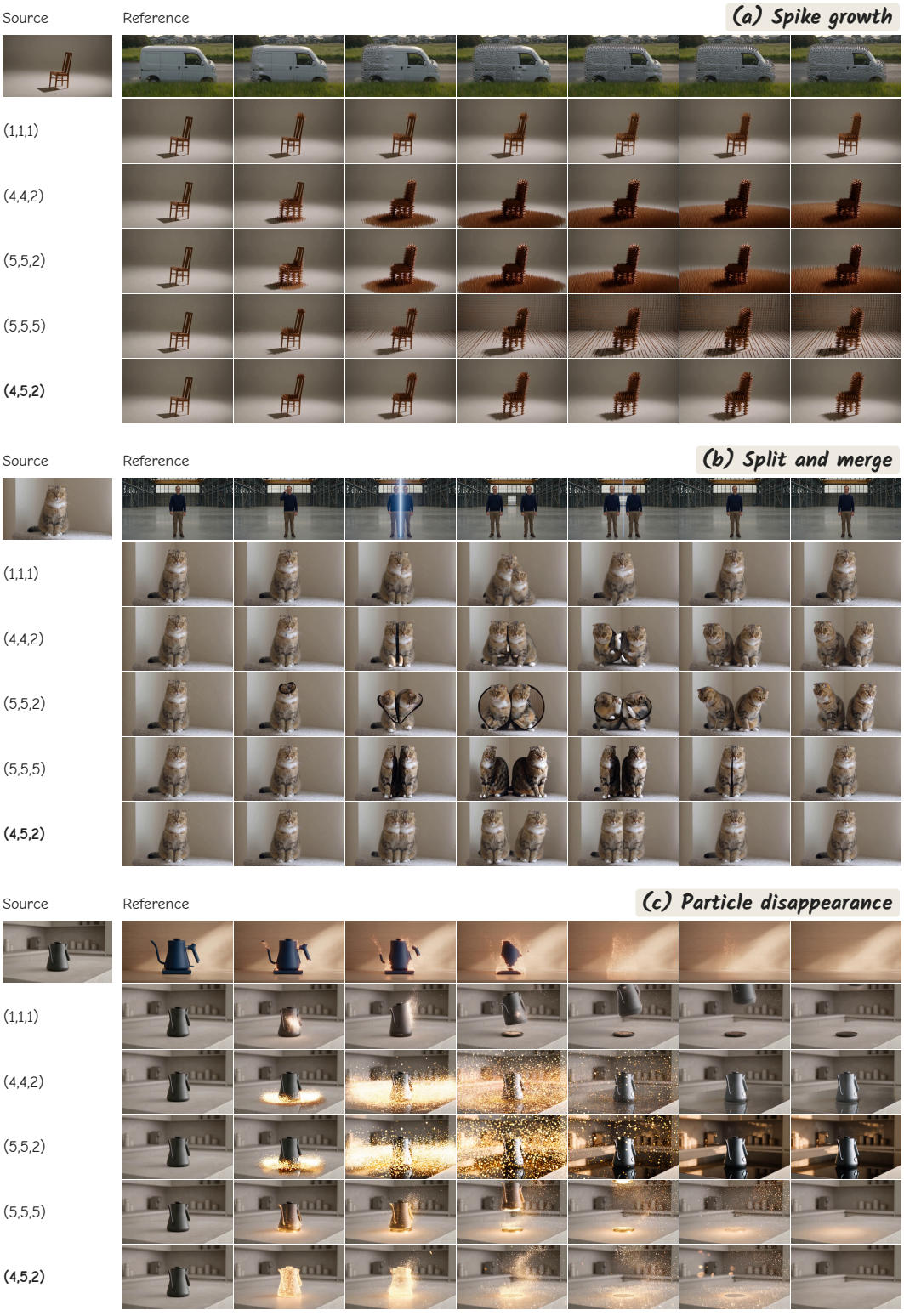

Figure 11: Selected comparisons of five guidance settings with matched inputs and sampling seeds. Rows show the same seven output frame indices; $(4, 5, 2)$ is our default.

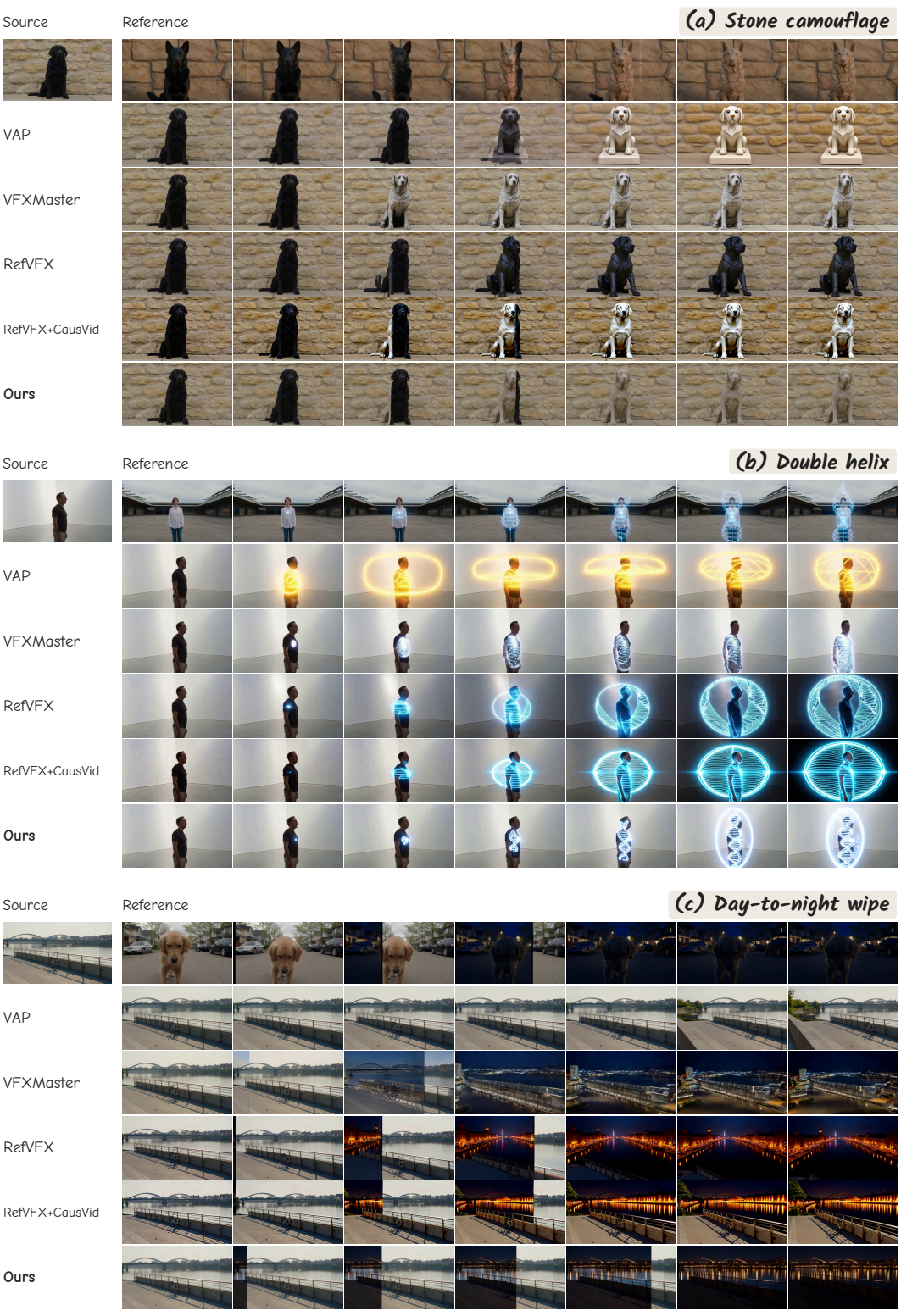

Figure 12: Selected baseline comparisons for stone camouflage, double-helix formation, and day-to-night transitions. All five methods use the same seven output frame indices; reference frames span each demonstration.

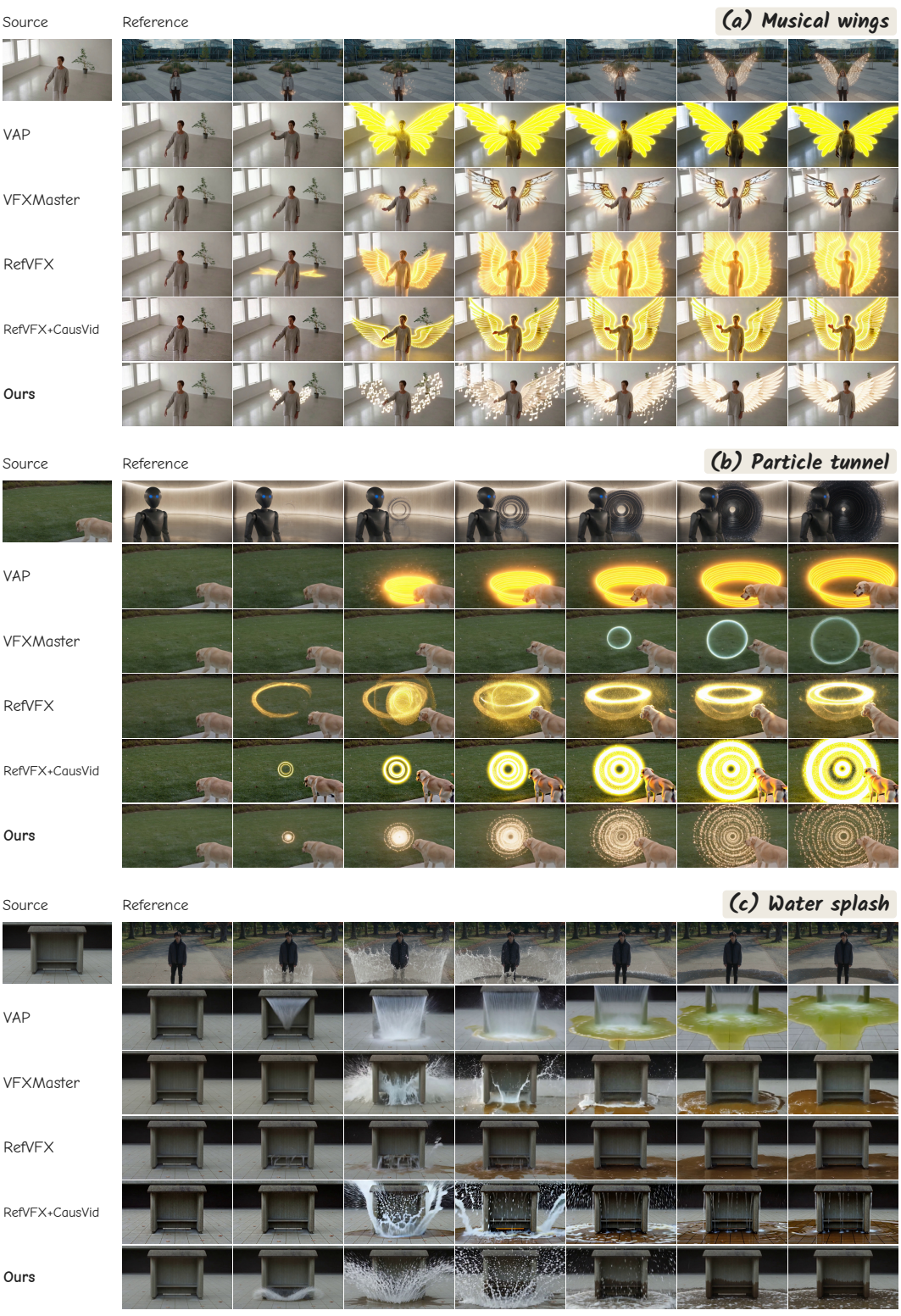

Figure 13: Selected baseline comparisons for musical wings, particle tunnels, and water splashes. The sampling protocol matches Figure 12.

Table 12: Additional comparisons on 500 EffectData cases. All methods use the same cases. Higher scores and lower Avg. Rank are better; best and second-best displayed values are bold and underlined.

		GPT-5.6 Sol (0–5)						
Method	Steps	EF	SP	TC	RD	VQ	EL	Avg. Rank↓
VAP	50	2.432	3.094	2.828	<u>4.990</u>	2.572	2.784	6.92
VFXMaster	50	2.962	3.182	2.878	4.982	2.468	3.128	5.58
RefVFX	50	2.900	3.268	3.138	4.984	2.840	3.104	4.33
RefVFX+CausVid	6	3.030	3.264	<u>3.344</u>	4.986	<u>2.986</u>	3.126	3.33
Reveris w/o prior	50	2.776	3.158	3.022	4.966	2.616	2.930	6.00
Reveris w/o gate	50	2.582	3.122	2.908	4.982	2.572	2.864	6.67
Reveris Teacher	50	<u>3.070</u>	**3.444**	<u>3.344</u>	**4.992**	<u>2.986</u>	<u>3.320</u>	<u>1.92</u>
Reveris (Ours)	4	**3.176**	<u>3.428</u>	**3.370**	**4.992**	**3.066**	**3.392**	**1.25**

	Agent (0–1)				VBench (0–1)				
Method	SF	SRD	TC	RL	SC	BC	MS	AQ	IQ
VAP	0.680	0.896	0.792	0.939	0.826	0.886	0.984	0.549	0.581
VFXMaster	0.812	0.909	<u>0.857</u>	0.937	**0.867**	0.895	**0.987**	0.549	0.608
RefVFX	0.785	<u>0.913</u>	0.822	0.939	0.839	0.889	0.984	0.558	0.633
RefVFX+CausVid	0.820	**0.923**	**0.867**	<u>0.942</u>	**0.867**	<u>0.898</u>	0.984	**0.567**	**0.696**
Reveris w/o prior	0.772	0.905	0.784	0.937	0.813	0.880	0.976	0.556	0.623
Reveris w/o gate	0.708	0.898	0.801	<u>0.942</u>	0.823	0.884	0.982	0.558	0.618
Reveris Teacher	<u>0.822</u>	**0.923**	0.825	<u>0.942</u>	0.840	0.894	0.980	0.558	0.632
Reveris (Ours)	**0.832**	**0.923**	0.846	**0.943**	<u>0.853</u>	**0.901**	<u>0.985</u>	<u>0.564</u>	<u>0.640</u>

D Additional Evaluation on EffectData

Setup. We supplement the main experiments with 500 cases from EffectData (Yang et al., 2026). Each case pairs the first frame of one video with a different reference video from the same effect category, using separate target and reference descriptions. All eight methods receive the same case inputs through their respective conditioning interfaces, and produce 49-frame videos with the sampling configurations in Section C.1. All methods use the same descriptions and are evaluated along the same dimensions as in the main comparison.

Quantitative comparison. Reveris achieves EF 3.176, SP 3.428, and EL 3.392, exceeding RefVFX+CausVid by 0.146, 0.164, and 0.266, respectively (Table 12). VQ also improves (3.066 versus 2.986), while TC remains close (3.370 versus 3.344). Reveris has the best Avg. Rank at 1.25 and leads agent SF, whereas RefVFX+CausVid obtains higher agent TC (0.867 versus 0.846) and VBench imaging quality. Compared with its teacher, the four-step student raises EF by 0.106 and VQ by 0.080; SP is 0.016 lower and RD is unchanged.

Qualitative comparison. Figure 14 compares feathered wings, electromagnetic charging, and a swirling portal. Reveris unfolds layered feathers while retaining the source portrait's lighting and clothing. For weapon charging, it forms a single blade with electric arcs; RefVFX produces an electric band, and RefVFX+CausVid produces crossed blades. In the portal example, Reveris develops a textured spiral behind the subject, while VFXMaster produces lightning and RefVFX forms diffuse storm clouds.

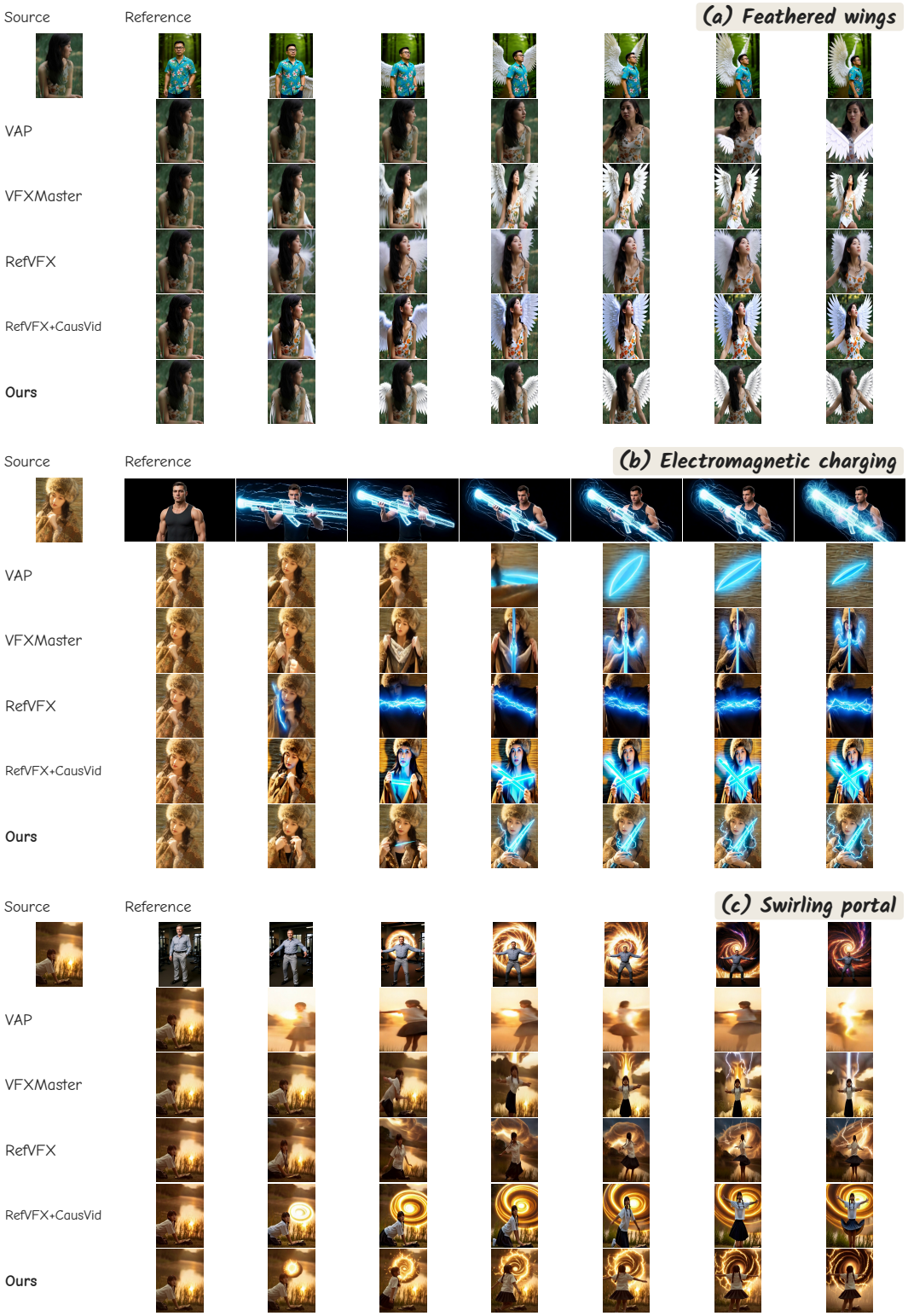

Figure 14: Selected comparisons on EffectData: feathered wings, electromagnetic charging, and a swirling portal. All methods use the same seven output frame indices; reference frames span each demonstration.