# HUMANFX: HUMAN-CENTRIC VISUAL EFFECTS GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Visual effects (VFX) combine exaggerated changes in appearance with nonphysical dynamics. Generating these effects for human subjects requires preserving facial identity, body structure, and motion continuity as the effect develops. Multi-person scenes further require accurate effect assignment and coordinated interactions. We present HumanFX, an image-to-video framework for human-centric visual effects generation. First, identity-preserving supervised fine-tuning (SFT) strengthens facial conditioning through a dedicated reference branch. Face crops sampled from video frames provide detailed identity information across varied views and expressions. Second, human-centric reinforcement learning (RL) uses anatomical and motion rewards to improve body structure and temporal consistency, with supervised regularization to retain expressive effect dynamics. Finally, because high-quality multi-person effect data are scarce, we introduce vision-language model (VLM)-guided test-time optimization (TTO). A differentiable visual question-answering (VQA) reward guides updates to the initial noise, refining effect assignment and interaction within a small computational budget. Extensive experiments demonstrate improvements in identity preservation, human motion, effect alignment, and user preference, alongside efficient multi-person control.

## 1 INTRODUCTION

Visual effects (VFX) extend video beyond captured reality through stylized appearances, exaggerated transformations, and nonphysical dynamics. Creating these effects traditionally requires specialized expertise and substantial manual effort (Liu et al., 2025b; Yang et al., 2026). Recent video generation methods learn effects through dedicated adapters (Liu et al., 2025b), combine effect experts (Mao et al., 2026), or transfer visual patterns from reference videos (Li et al., 2026; Bian et al., 2026; Jones et al., 2026; Yang et al., 2026). An image-to-video interface lets users create effects by providing an image and describing the desired result in natural language. This makes effect creation accessible without professional editing skills.

Human-centric VFX must reproduce the requested effect while keeping the person recognizable and their movement coherent. When conditioning relies only on the first frame, facial information is limited to one viewpoint. The face may also occupy a small image region, leaving fine identity features poorly represented. These limitations become more pronounced when the subject turns, moves rapidly, or reappears after an occlusion (Xue et al., 2026). Maintaining a well-formed body and coherent movement poses a further challenge. Extra limbs, jitter, and abrupt orientation changes can occur even when the overall effect looks convincing (Wang et al., 2024; Fang et al., 2026). Vision-language models (VLMs) assess high-level semantic content, but can miss these local and short-lived defects (Wang et al., 2024; Hong et al., 2025; Thomas et al., 2025). Assessing more frames with a VLM increases computational cost, while sparse sampling can miss brief motion errors (Hong et al., 2025). These failures call for supervision that measures human anatomy and motion as the effect develops.

Multi-person VFX additionally require control over effect assignment and interaction. A prompt may specify an effect for one person and reactions from the others, or request coordinated effects across several people. For example, "the center person rises into the sky while the others look up in surprise" requires both the intended target and the requested reactions. Applying the effect to another person or omitting the reactions violates the instruction. High-quality training videos covering these



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.

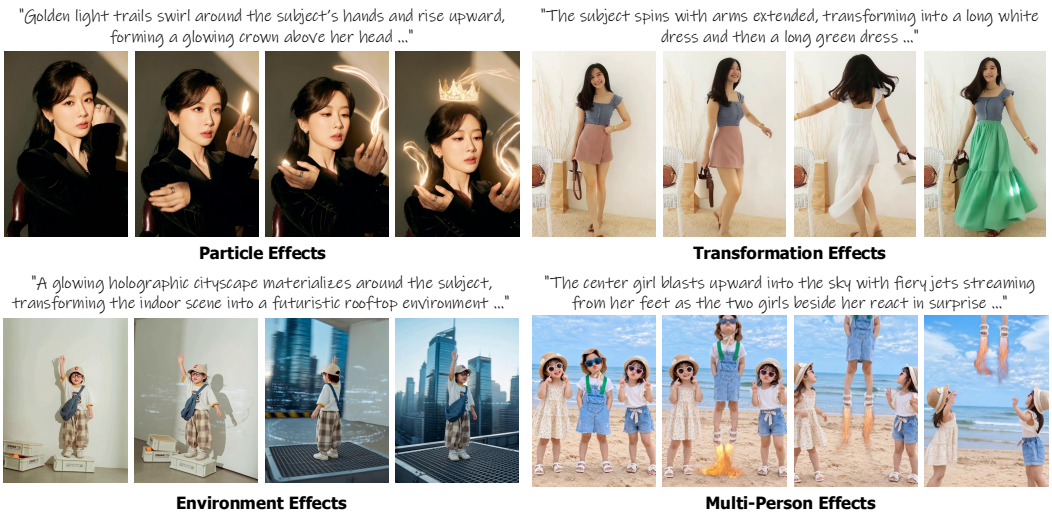


Figure 1: **HumanFX generates diverse human-centric visual effects.** Examples include particle effects, appearance transformations, environmental changes, and multi-person scenes.

combinations are scarce. EffectData (Yang et al., 2026), for example, excludes multi-person images during subject collection. Single-person examples demonstrate effect appearance and dynamics, but provide no supervision for assigning effects and coordinating reactions across several people.

We introduce HumanFX, a framework for human-centric image-to-video effects generation. We train the model on single-person VFX data and refine effect assignment and interaction for multi-person inputs at test time. Training combines identity-preserving supervised fine-tuning (SFT) with human-centric reinforcement learning (RL). During SFT, a dedicated identity branch receives a face crop sampled from a video frame in which the subject's face is visible. The crop emphasizes facial detail, and sampling across video frames supplies varied views and expressions. Human-centric RL then targets challenging effects involving complex movements, using anatomical and motion rewards to reduce structural defects and temporal discontinuities. Supervised regularization preserves effect appearance and expressive dynamics during reward optimization.

Since high-quality multi-person effect data are limited, we extend the trained model through efficient VLM-guided test-time optimization (TTO). The VLM evaluates which people receive each effect and whether the requested actions and interactions occur. A differentiable visual question-answering (VQA) reward turns this feedback into gradients for the initial noise while the model parameters remain fixed. A small number of guided updates can thus refine the current prediction toward the instruction. Best-of-$N$ uses the reward only to select among independently sampled candidates. Our method instead uses the reward gradient to guide the next update, making more effective use of feedback within a limited inference budget.

Our contributions are threefold. We introduce identity-preserving SFT with facial references that provide detailed identity information across varied video frames. We develop human-centric RL for challenging effects, combining anatomical and motion rewards with supervised regularization for effect fidelity. We further introduce VLM-guided TTO for multi-person effect assignment and interaction without additional multi-person training videos. Experiments show that focused post-training and test-time feedback enable a compact generator to preserve facial identity and coherent movement while improving effect instruction following, scene continuity, and user preference. Figure 1 illustrates diverse visual effects supported by HumanFX.

## 2 Related Work

**Visual effects generation.** Building on latent diffusion (Rombach et al., 2022) and transformer architectures (Peebles & Xie, 2023), video generators learn broad visual and motion priors (Yang et al., 2025; Team Wan et al., 2025). MagicVFX (Guo et al., 2024a) and VFX Creator (Liu et al., 2025b) adapt these priors to individual effects, while Omni-Effects (Mao et al., 2026) combines

effect experts with spatial control. VFXMaster (Li et al., 2026), Video-As-Prompt (Bian et al., 2026), RefVFX (Jones et al., 2026), and EffectMaker (Yang et al., 2026) transfer effect appearance and dynamics from reference videos. Identity-preserving video generation uses facial representations (He et al., 2024; Yuan et al., 2025) and reference attention (Xue et al., 2026). However, effect and identity conditioning leave anatomical and motion errors unresolved. We combine identity-preserving SFT with human rewards and use VLM feedback to refine multi-person effect assignment and interaction.

**Reinforcement learning for video generation.** Reward-based diffusion training uses scalar feedback (Black et al., 2024; Zheng et al., 2026) or gradients from differentiable evaluators (Clark et al., 2024; Prabhudesai et al., 2024). VideoReward (Liu et al., 2025a) learns rewards for visual quality, motion, and text alignment, while differentiable VLM rewards provide semantic supervision (Wang et al., 2026). However, fine anatomical defects and brief motion errors remain difficult for general-purpose VLM judges to assess (Wang et al., 2024; Hong et al., 2025; Thomas et al., 2025). We use dedicated anatomical and motion rewards to improve human consistency in challenging effects. Supervised regularization (Ye et al., 2025) preserves the appearance and dynamics of the effect during this optimization.

**Test-time optimization for video generation.** Test-time scaling improves generation through candidate selection (Ma et al., 2025) or optimization of the initial noise (Guo et al., 2024b; Eyring et al., 2024). Diffusion-DRF (Wang et al., 2026) derives differentiable rewards from VLM responses, and VLM-as-Teacher (Cheng et al., 2026) uses visual feedback for online LoRA adaptation. These approaches provide semantic feedback beyond the training examples. For VFX, limited multi-person data leave effect assignment and interaction insufficiently supervised. We address this gap with a differentiable VQA reward that refines the initial noise for each input while keeping the generator fixed.

# 3 Method

## 3.1 Overall Framework

Given an image $I$ and a text description $c$, HumanFX generates a video $\mathbf{v}$ with the requested effects and human behavior. The image establishes the scene and its subjects, while the description specifies how they transform, move, and interact. Figure 2 presents our three-stage framework. Identity-preserving SFT learns effect appearance and dynamics from single-person videos while retaining facial details. Human-centric RL improves anatomy and motion, with supervised regularization preserving the intended effect. For multi-person inputs, VLM-guided TTO refines effect assignment and interaction through a differentiable VQA reward. It updates the initial noise for the current instruction while keeping the trained generator fixed.

## 3.2 Identity-Preserving Supervised Fine-Tuning

Facial identity must remain recognizable as the effect changes the subject's appearance and movement. The first frame provides one facial view, often at a scale that obscures fine facial detail. We augment first-frame conditioning with an identity reference $r$, obtained by cropping the subject's face from a randomly selected video frame in which it is visible. The crop emphasizes facial features and proportions that are difficult to represent at the scale of the full scene. Sampling across video frames also exposes the model to different views and expressions of the same person. Together, detailed facial encoding and varied reference views provide identity information that remains useful through changes in pose, appearance, and visibility.

We encode the face crop and video with the VAE and process their features in identity and video branches with separate low-rank adaptation (LoRA) modules (Hu et al., 2022). The video LoRA learns effect appearance and dynamics, while the identity LoRA learns facial conditioning through separate query, key, and value projections. Restricted attention keeps reference features independent of the evolving video content. Video queries attend to both streams, whereas reference queries attend only to the identity stream. For video features $(Q_v, K_v, V_v)$ and reference features $(Q_r, K_r, V_r)$, the attention outputs are

$$O_v = \mathrm{Attn}(Q_v, [K_v; K_r], [V_v; V_r]), \qquad O_r = \mathrm{Attn}(Q_r, K_r, V_r). \tag{1}$$

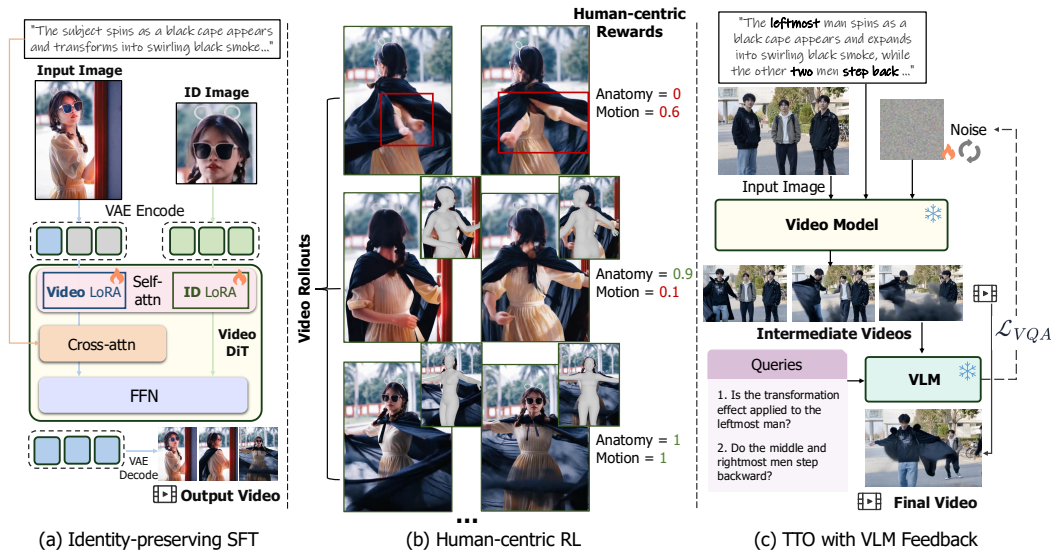


Figure 2: **Overview of HumanFX.** (a) Identity-preserving supervised fine-tuning learns effects with detailed facial references. (b) Human-centric reinforcement learning improves anatomy and motion while supervised regularization retains effect dynamics. (c) For multi-person inputs, VLM-guided TTO evaluates effect assignment and interaction and refines the initial noise using VQA gradients.

Here $[\cdot;\cdot]$ denotes concatenation along the sequence dimension. Separate positional coordinates distinguish the face image from the video sequence.

This asymmetric attention allows the video branch to retrieve facial information throughout generation while the identity branch retains the reference representation. During supervised training, the complete target video teaches how the effect develops, and the sampled face crop provides a complementary view of the subject. This cue is useful when an effect involves turning or temporary occlusion. At inference, $r$ can be another image of the same person. Section 4.1 describes our alternate-view reference preparation, and Appendix A gives the attention and training details.

### 3.3 Human-Centric Reinforcement Learning

Identity-preserving SFT learns effect dynamics with stronger facial conditioning, but complex movements can still produce malformed limbs and unstable motion. An extra hand is a local spatial defect, whereas jitter or an abrupt orientation change becomes apparent across neighboring frames. General-purpose VLM judges can miss these spatial and temporal details, and evaluating more frames increases their computational cost (Wang et al., 2024; Hong et al., 2025). We use specialized anatomical detection and human motion recovery to evaluate the generated person (Fang et al., 2026). These complementary measurements guide RL on challenging single-person effects involving movements such as turning, spinning, dancing, and running, where human structure and motion continuity are most difficult to maintain.

VFX can deliberately change appearance and depart from ordinary motion. Our rewards target unintended structural and temporal artifacts in intervals where the subject remains visible in human form. The anatomical reward evaluates the body's structure within each such interval. A human artifact detector (Wang et al., 2024) localizes extra arms, hands, legs, and feet. The reward $R_{\mathrm{a}}$ decreases with the severity, frequency, and persistence of detected artifacts. Motion assessment complements this spatial evidence since individually well-formed poses can still form an unstable sequence. We recover human motion in gravity-aligned world coordinates (Shen et al., 2024) to separate body movement from camera motion. A smoothness score $R_s$ penalizes excessive angular acceleration and jerk, and an orientation score $R_o$ captures abrupt changes in body or head heading. The combined reward is

$$R_{\mathrm{human}} = \lambda_{\mathrm{a}} R_{\mathrm{a}} + \lambda_{\mathrm{m}} R_{\mathrm{m}}, \qquad R_{\mathrm{m}} = \frac{R_s + R_o}{2}, \qquad \lambda_{\mathrm{a}} + \lambda_{\mathrm{m}} = 1. \tag{2}$$

Anatomical detection identifies local structural defects, while motion recovery identifies discontinuities in the pose sequence. Intended disappearance and nonhuman transformations are excluded from both rewards.

Human rewards must improve the execution of an effect without suppressing the movement it requires. Optimizing these rewards alone can favor limited movement, making anatomical and temporal defects easier to avoid. We therefore combine reward optimization with supervised regularization on ground-truth videos from the effect dataset. The generated samples expose anatomical and motion defects, while the supervised targets retain the appearance changes and expressive dynamics that define the effect. Using DiffusionNFT (Zheng et al., 2026), we optimize

$$\mathcal{L}_{\mathrm{RL}} = \mathcal{L}_{\mathrm{NFT}} + \lambda_{\mathrm{SFT}} \mathcal{L}_{\mathrm{SFT}}. \tag{3}$$

The two objectives encourage coherent human motion while retaining the appearance changes and movement required by the effect (Ye et al., 2025). We update the video LoRA and keep the identity adapter fixed to preserve its facial conditioning. Appendix B defines the rewards and training objective.

### 3.4 VLM-Guided Test-Time Optimization

The SFT and RL stages learn effect appearance, identity preservation, and coherent human motion from single-person videos. These examples associate an effect with one subject, but do not specify how to distribute it among several people or coordinate their reactions. High-quality training videos covering these combinations are scarce. We use the VLM's semantic understanding to evaluate effect assignment and interaction for each input, then convert this feedback into a differentiable VQA reward. This provides supervision for the requested event without changing the trained generator.

The reward must assess the effect and the participants' behavior together. For example, "the leftmost man turns into smoke while the other two step back" specifies a transformation, its recipient, and the reactions of two other people. Detecting smoke alone would overlook a mistaken recipient or missing reactions. Given the input image and description, the VLM proposes focused questions that assess these requirements, with "Yes" as the desired answer (Cheng et al., 2026; Wang et al., 2026). The questions check whether the effect reaches the designated people and whether the requested actions, reactions, and interactions occur. Let $\mathcal{Q}$ denote the question set and $a^+$ the token sequence for "Yes". For an intermediate video prediction $\tilde{\mathbf{v}}$, the frozen VLM $P_\psi$ defines

$$\mathcal{L}_{\mathrm{VQA}}(\tilde{\mathbf{v}}) = -\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \sum_{j=1}^{|a^+|} \log P_\psi(a_j^+ \mid I, c, \tilde{\mathbf{v}}, q, a_{<j}^+). \tag{4}$$

A lower loss indicates greater VLM confidence that the requested effect and interactions occur. Its gradient guides refinement of the initial noise. By evaluating the participants' roles alongside the effect, it captures semantic errors that appearance-based assessment alone would miss.

To make repeated assessment efficient, the RL-trained model is distilled into a four-step generator following DMD2 (Yin et al., 2024). Starting from initial noise $\mathbf{z}$, the generator estimates a clean video latent $\hat{\mathbf{z}}_0(\mathbf{z})$ during denoising. A lightweight surrogate decoder $\mathcal{D}_{\mathrm{lite}}$ from LightX2V (LightX2V Contributors, 2025) converts this estimate into an intermediate video prediction. The VQA loss is differentiable through this prediction, so its gradient can update the noise that determines the generated event. For an initial sample $\mathbf{z}^{(0)} \sim \mathcal{N}(0, \mathbf{I})$, we solve

$$\mathbf{z}^\star = \arg \min_{\|\mathbf{z}\|_2 = \|\mathbf{z}^{(0)}\|_2} \mathcal{L}_{\mathrm{VQA}}\big(\tilde{\mathbf{v}}(\mathbf{z})\big), \qquad \tilde{\mathbf{v}}(\mathbf{z}) = \mathcal{D}_{\mathrm{lite}}\big(\hat{\mathbf{z}}_0(\mathbf{z})\big). \tag{5}$$

We approximately solve this objective with a small number of spherical updates that preserve the initial noise norm. Gradients pass through the frozen VLM, lightweight decoder, and generator, while only the noise changes. Optimization stops when the VQA loss reaches a threshold or the update budget is exhausted. The refined noise then initializes final generation with the four-step model and standard decoder. Appendix C gives the clean-latent estimate, question-generation prompt, and optimization settings.

The key advantage over best-of-$N$ is how feedback guides the search. Candidate selection uses a reward to rank independent noise samples. Instead, our method uses the reward gradient to improve

the current sample, guiding successive updates toward the requested assignment and interaction. Intermediate predictions lower the cost of assessment, but the gradient supplies the direction of refinement. This lets the generator respond to a new arrangement of people and their requested roles while retaining the effect and human-preservation capabilities learned from single-person videos.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Implementation details.** We use Wan2.2-TI2V-5B (Wan Team, 2025) with video and identity LoRA adapters of ranks 256 and 128. SFT runs for 4,000 steps at a learning rate of $5 \times 10^{-5}$. DiffusionNFT (Zheng et al., 2026) runs for 50 steps at $10^{-5}$, with 16 rollouts per input condition and supervised regularization. Training uses a 720p pixel budget and a batch size of 32 on 32 NVIDIA H20 GPUs. At inference, Qwen-Image-Edit (Wu et al., 2025) provides an alternate facial view for single-person inputs. Multi-person generation uses the first frame without the identity branch. The RL-trained model is distilled into a four-step generator following DMD2 (Yin et al., 2024). Qwen3-VL-4B (Bai et al., 2025) provides the VQA reward for at most four noise updates. Outputs contain 121 frames at 24 fps. Appendix D.1 gives the remaining settings.

**Datasets.** We construct HumanFX-100K by filtering single-person effect clips from Open-VFX (Liu et al., 2025b), VAP-Data (Bian et al., 2026), and EffectData (Yang et al., 2026). The resulting 100K-video corpus covers diverse effects, such as environmental effects, transformations, stylization, camera effects, and composite effects. SFT uses the full corpus. For RL, we select challenging single-person effects with large or complex movements, such as turning, spinning, dancing, and running. Supervised regularization continues to sample from the full corpus. Following the Open-VFX protocol used by EffectMaker (Yang et al., 2026), we evaluate 14 effect categories with 10 held-out human images per category, giving 140 image–prompt pairs with effect instructions. The same set is used for controlled ablations. We additionally construct 60 multi-person pairs that specify effect recipients and associated actions, reactions, or interactions. Evaluation identities, source images, and derived videos are excluded from training and reward calibration. We generate three outputs per input. Appendix E details the evaluation protocol.

**Baselines.** We compare with Wan2.2 (Wan Team, 2025), VFX Creator (Liu et al., 2025b), Omni-Effects (Mao et al., 2026), and MiniMax-H3 (MiniMax, 2026a). All receive the same initial image and effect description. VFX Creator uses the corresponding effect adapter and required spatial condition from the input image. Omni-Effects uses its released image-to-video checkpoint. The Wan2.2 SFT ablation fine-tunes the same backbone on our effect corpus without the identity branch or human rewards.

**Evaluation metrics.** We evaluate visual quality (VQ) and motion quality (MQ) with VideoReward (Liu et al., 2025a), following EffectMaker (Yang et al., 2026). Following Stand-In (Xue et al., 2026), face similarity (FaceSim) averages cosine similarities between CurricularFace (Huang et al., 2020) embeddings of the input face and those of detected faces in the generated video frames. Following Video-As-Prompt (Bian et al., 2026), we report VBench (Huang et al., 2024) motion smoothness (Smooth.) and dynamic degree (Dyn.). These metrics distinguish temporal smoothness from the presence of motion. GPT-6 Astra (OpenAI, 2026) evaluates effect alignment (Align.) from the generated video and instruction using the assessment rules adapted from Video-As-Prompt. User preference (Pref.) is each method's share of selections among five results for the same input. Twenty participants, six with experience in design or film and video production, each evaluate the same 72 inputs. We also provide qualitative comparisons to assess the overall quality of the generated effects. Appendix E.2 details the evaluation procedures.

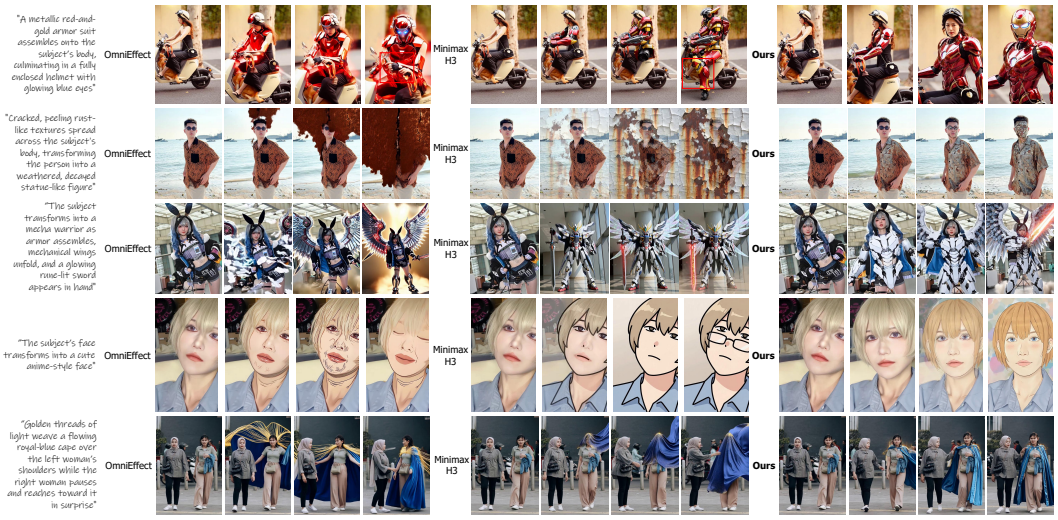


Figure 3: **Visual comparison across diverse effects.** HumanFX generates the requested effects while preserving recognizable subjects and coherent human structure.

### 4.2 Comparison with Image-to-Video Methods

Table 1: **Comparison with image-to-video methods.** VQ and MQ denote visual and motion quality. FaceSim is face similarity, Smooth. is motion smoothness, Dyn. is dynamic degree, Align. is effect alignment, and Pref. is user preference. Bold and underlined values mark the best and second-best results for metrics with arrows. Dyn. and Pref. are percentages.

| | Quality | | Human | Motion | | Effect generation | |
|---|---|---|---|---|---|---|---|
| Method | VQ↑ | MQ↑ | FaceSim↑ | Smooth.↑ | Dyn.↑ | Align.↑ | Pref.↑ |
| Wan2.2 | 2.18 | 0.11 | 0.634 | 97.7 | 69.5 | 30.2 | 1.3 |
| VFX Creator | 2.13 | -0.06 | 0.646 | 97.9 | 72.1 | 66.7 | 4.8 |
| Omni-Effects | 2.32 | -0.12 | 0.652 | 98.0 | 71.7 | 68.3 | 5.3 |
| MiniMax-H3 | **3.18** | **0.31** | <u>0.711</u> | **98.5** | <u>73.3</u> | <u>72.2</u> | <u>39.6</u> |
| HumanFX | <u>3.14</u> | <u>0.28</u> | **0.737** | <u>98.2</u> | **77.6** | **75.8** | **49.0** |

**Human identity preservation and motion quality.** HumanFX achieves the highest FaceSim in Table 1, outperforming MiniMax-H3, VFX Creator, and Omni-Effects. As illustrated in Figure 3, these quantitative gains are also reflected in the qualitative results. In the fourth row, MiniMax-H3 alters the subject's facial appearance and introduces glasses during stylization, while HumanFX retains recognizable facial features. In the first row, MiniMax-H3 produces malformed legs during armor assembly, as highlighted by the red box, while HumanFX preserves coherent human anatomy. These examples demonstrate the complementary roles of the face-reference branch and human-centric rewards in preserving identity and human structure under challenging visual effects.

HumanFX also achieves a higher dynamic degree than MiniMax-H3, with more expressive movement in the qualitative results. MiniMax-H3's slightly higher motion smoothness accompanies more restrained movement and sometimes incomplete effect execution. Smoothness alone can favor nearly static results (Liao et al., 2024), so it must be considered alongside dynamic degree and effect alignment. Our human rewards improve anatomical and temporal consistency, while supervised regularization preserves the dynamics required by the effect.

**Effect fidelity and visual quality.** MiniMax-H3 uses a 33B pretrained backbone (MiniMax, 2026b) and retains a slight lead in VQ and MQ. HumanFX builds on the smaller Wan2.2 backbone, yet focused post-training yields the highest effect alignment, with stronger dynamics and identity

Table 2: **Core training ablations.** The first two rows compare SFT variants. The remaining rows compare RL variants. Rewards $R_a$ and $R_m$ are scaled by 100.

| Variant | FaceSim↑ | $R_a$ ↑ | $R_m$ ↑ | Dyn.↑ | Align.↑ |
|---|---|---|---|---|---|
| Wan2.2 SFT | 0.709 | 82.7 | 80.1 | **77.6** | 75.4 |
| + Identity branch | 0.731 | 82.2 | 79.6 | 77.1 | 75.2 |
| w/o anatomical reward | 0.735 | 84.5 | 90.8 | 76.9 | **75.8** |
| w/o motion reward | 0.734 | 92.0 | 82.7 | 77.4 | 75.4 |
| w/o SFT loss | 0.735 | **94.2** | **94.8** | 71.2 | 67.6 |
| HumanFX | **0.737** | 92.8 | 91.3 | **77.6** | **75.8** |

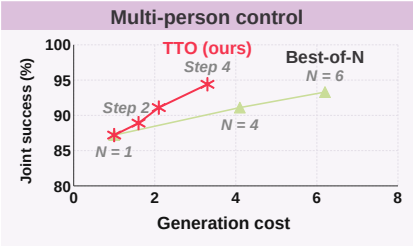


Figure 4: **Multi-person control versus cost.** Generation cost is reported as relative runtime to compare inference overhead, with one four-step generation set to 1. Steps denote the maximum number of noise updates.

preservation. VFX Creator and Omni-Effects improve effect execution over the general-purpose backbone, but remain weaker in visual quality and instruction following.

Figure 3 shows that successful effect generation requires integrating the visual change with the subject. In the second row, MiniMax-H3 spreads the rust texture across the scene and obscures the person, while HumanFX applies it to the body and preserves the surrounding scene. In the fifth row, MiniMax-H3's cape covers the person's head and upper body, while HumanFX forms a cape that rests naturally on the shoulders. In both examples, our effects remain more naturally integrated with the subject.

**User preference.** HumanFX receives $49.0\%$ of selections, compared with $39.6\%$ for MiniMax-H3. This preference is consistent with the more coherent effect development in the qualitative comparisons. In the third row of Figure 3, MiniMax-H3 abruptly cuts to the armored character, interrupting the transition from the input subject. HumanFX develops the armor and wings continuously while keeping the subject recognizable. Such continuity complements the identity preservation and natural effect integration shown in the other rows. The preference results favor these qualities of the complete video over MiniMax-H3's slight advantage in VQ and MQ.

These results show that focused post-training enables a compact generator to achieve stronger identity preservation, effect execution, and user preference. Additional qualitative results are provided in the supplementary video material.

### 4.3 Ablation Study

We study how identity conditioning, human rewards, supervised regularization, and test-time feedback contribute to effect generation. Table 2 evaluates the training components on single-person effects, and Figure 4 compares TTO with independent sampling on multi-person inputs. Figure 5 shows representative visual comparisons. The anatomical and motion rewards in Table 2 track the properties optimized during training, and the videos show how these changes affect the generated subjects.

**Identity preservation.** Wan2.2 SFT learns effect appearance and dynamics from the same training videos, but relies on the full input image for facial information. Adding the identity branch improves identity consistency while retaining the intended effect. The reference crop emphasizes facial detail, and sampling across video frames provides complementary views and expressions. These cues help retain the same person as pose and appearance change. The example without the identity branch in Figure 5 shows the resulting loss of facial consistency.

**Human structure and motion.** Removing either human reward weakens the corresponding aspect of the generated person. The visual comparisons reveal the distinction between structural and temporal errors. A malformed limb is visible within a frame, whereas an abrupt change in orientation disrupts the movement across frames. Individually plausible poses therefore do not guarantee a coherent sequence, and smooth movement does not guarantee well-formed anatomy. Combining anatomical and motion rewards addresses both sources of error. SFT learns from effect examples, while human-centric RL penalizes structural and temporal defects in the model's own generations.




Figure 5: **Visual ablations of HumanFX.** Wan2.2 SFT provides an effect-only training baseline. The other panels remove one component from HumanFX. Each upper sequence is paired with the full method below. The comparisons show the roles of identity conditioning, anatomical and motion rewards, supervised regularization, and TTO in preserving the person, generating expressive effects, and refining multi-person generation.

**Supervised regularization.** Without supervised regularization, higher human rewards coincide with weaker dynamics and less complete effect execution. Reward optimization can favor limited movement, making anatomical and temporal defects easier to avoid. Figure 5 illustrates the resulting loss of expressive movement. Supervised examples preserve the appearance changes and actions that define the effect, keeping human consistency tied to successful effect generation. Human rewards and supervised examples thus serve complementary roles: the rewards reduce defects, while the examples preserve the intended effect appearance and dynamics.

**Multi-person control.** On the multi-person set, VLM-guided TTO improves joint success from $87.2\%$ without refinement to $94.4\%$ with at most four updates. Joint success requires correct effect assignment together with the requested actions, reactions, and interactions. An effect can look convincing yet remain incorrect because it appears on the wrong person or the others do not respond as specified. The VQA objective evaluates these roles alongside the effect, providing the semantic feedback missing from single-person training. The multi-person example in Figure 5 compares generation with and without this refinement.

Figure 4 compares TTO update budgets with best-of-$N$ sampling (Ma et al., 2025). The budget of two TTO updates matches best-of-four sampling with roughly half the generation cost. Four updates reach $94.4\%$ joint success, compared with $93.3\%$ for best-of-six sampling, at substantially lower cost. The distinction is how each method uses feedback. TTO follows the VQA gradient to improve its current sample, while best-of-$N$ selects among independently sampled candidates. Both methods share the distilled generator and independent evaluator. Guided refinement therefore makes more effective use of a limited inference budget, improving assignment and interaction without collecting additional multi-person training videos. Appendix C.2 details the cost accounting.

## 5 Conclusion

We presented HumanFX for human-centric image-to-video effects generation. Identity-preserving SFT uses face crops sampled across video frames to provide detailed identity cues from varied views. Human-centric RL improves anatomy and motion, while supervised regularization retains expressive effect dynamics. VLM-guided TTO extends the model to multi-person inputs by refining the initial noise according to the requested assignment and interaction. Together, focused post-training and test-time feedback enable a compact generator to produce expressive effects with consistent identity and coherent human motion. Quantitative and qualitative comparisons show stronger instruction following, more continuous effects, and higher user preference, alongside efficient multi-person control.

### AI USE STATEMENT

Generative AI tools assisted with language editing to improve clarity and readability. As described in the paper, pretrained generative models and VLMs were also used for dataset captioning, face synthesis, test-time VQA feedback, and effect-alignment evaluation. All AI-assisted content was reviewed by the authors, who take full responsibility for the final content of this work.

### ETHICS STATEMENT

This work studies human-centric visual effects generation for academic research and creative applications. The image and video data come from publicly available datasets or synthetic generation and are used in accordance with their licenses and terms of use. We do not collect private images or videos. Participants in the human preference study provided informed consent before evaluation. No sensitive personal information was collected, and the study recorded only anonymous preference judgments. Identity-preserving generation can be misused for impersonation or deceptive content. Applications should obtain appropriate permission for likeness use and clearly identify generated content.

### REPRODUCIBILITY STATEMENT

Section 3 describes the model and objectives. Appendices A and B detail reference preparation and human rewards. The test-time algorithm, VLM prompts, and computational costs appear in Appendix C. Data preparation and training settings are given in Appendix D, and evaluation inputs and metrics in Appendix E. Upon acceptance, we will release the code.

## REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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 (ICLR)*, 2026.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Junhao Cheng, Liang Hou, Tianxiong Zhong, Xin Tao, Pengfei Wan, Kun Gai, and Jing Liao. VLMs are good teachers for video reasoning via adaptive test-time optimization. *arXiv preprint arXiv:2606.02564*, 2026.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5203–5212, 2020.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: Enhancing one-step text-to-image models through reward-based noise optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yusu Fang, Tiange Xiang, Tian Tan, Narayan Schuetz, Scott Delp, Li Fei-Fei, and Ehsan Adeli. HumanScore: Benchmarking human motions in generated videos. *arXiv preprint arXiv:2604.20157*, 2026.

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

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. InitNO: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9380–9389, 2024b.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. ID-Animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. MotionBench: Benchmarking and improving fine-grained video motion understanding for vision language models. *arXiv preprint arXiv:2501.02955*, 2025.

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

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. CurricularFace: Adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5901–5910, 2020.

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

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 (ECCV)*, 2026.

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, 2026.

Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. In *Advances in Neural Information Processing Systems*, 2024.

LightX2V Contributors. LightX2V: Light video generation inference framework, 2025. URL `https://github.com/ModelTC/LightX2V`.

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

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*, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2523–2534, 2025.

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*, 2026.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities, 2026a. URL `https://www.minimax.io/blog/minimax-h3`. Released July 31, 2026.

MiniMax. MiniMax-H3 model card, 2026b. URL https://huggingface.co/MiniMaxAI/MiniMax-H3. Accessed September 24, 2026.

OpenAI. GPT-6 Astra model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Accessed September 25, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Qwen Team. Qwen3.5-122B-A10B model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-122B-A10B.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

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

Xavier Thomas, Youngsun Lim, Ananya Srinivasan, Audrey Zheng, and Deepti Ghadiyaram. Generative action tell-tales: Assessing human motion in synthesized videos. *arXiv preprint arXiv:2512.01803*, 2025.

Wan Team. Wan2.2-TI2V-5B model card, 2025. URL https://huggingface.co/Wan-AI/Wan2.2-TI2V-5B.

Kaihong Wang, Lingzhi Zhang, and Jianming Zhang. Detecting human artifacts from text-to-image models. *arXiv preprint arXiv:2411.13842*, 2024.

Yifan Wang, Yanyu Li, Gordon Guocheng Qian, Sergey Tulyakov, Yun Fu, and Anil Kag. Diffusion-DRF: Free, rich, and differentiable reward for video diffusion fine-tuning. *arXiv preprint arXiv:2601.04153*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Bowen Xue, Zheng-Peng Duan, Qixin Yan, Wenjing Wang, Hao Liu, Chun-Le Guo, Chongyi Li, Chen Li, and Jing Lyu. Stand-In: A lightweight and plug-and-play identity control for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23314–23324, 2026.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025.

Haotian Ye, Kaiwen Zheng, Jiashu Xu, Puheng Li, Huayu Chen, Jiaqi Han, Sheng Liu, Qinsheng Zhang, Hanzi Mao, Zekun Hao, Prithvijit Chattopadhyay, Dinghao Yang, Liang Feng, Maosheng Liao, Junjie Bai, Ming-Yu Liu, James Zou, and Stefano Ermon. Data-regularized reinforcement learning for diffusion models at scale. *arXiv preprint arXiv:2512.04332*, 2025.

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 (NeurIPS)*, 2024.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 325–341, 2018.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988, 2025.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations (ICLR)*, 2026.

## A Identity-Preserving Supervised Fine-Tuning

**Training references.** For each single-person training video, we sample a frame in which the subject's face is visible. RetinaFace (Deng et al., 2020) detects and aligns the face to produce a $512 \times 512$ crop. Following Stand-In (Xue et al., 2026), we use BiSeNet (Yu et al., 2018) to prepare a version with a white background. We choose between the original and background-removed crops with equal probability and drop the reference with probability $0.2$.

**Reference attention.** The Wan2.2 VAE and patch embedding convert the face reference into 256 tokens. We assign them temporal coordinate $-1$ and spatial coordinates $(h + H_v, w + W_v)$, where $(h, w)$ indexes the reference grid and $(H_v, W_v)$ gives the video token grid dimensions. Rotary embeddings are applied before the attention operation in Eq. 1. The reference stream uses modulation at the clean timestep and shares the pretrained output and feed-forward projections. Its query, key, and value projections receive separate identity LoRA updates. Text cross-attention and the video LoRA operate on the video stream. Reference keys and values remain fixed across denoising steps and can be cached during single-person inference.

**Training.** We train both LoRA adapters on the single-person effect corpus with the standard flow-matching objective over generated latent coordinates. The first-frame latent remains clean, and the backbone, text encoder, and VAE remain fixed. Table 3 lists the adapter ranks and training settings.

## B Human Reward Definitions

**Reward calibration.** We manually annotate anatomical and motion defects in 200 videos, identifying the affected frames and intervals. We use these annotations to set detector thresholds, motion reference scales, and reward aggregation weights. The calibration set is separate from evaluation, and the calibrated parameters remain fixed across experiments. We score anatomy where the subject is visible in human form and motion within continuous visible intervals of at least eight frames.

**Anatomical severity.** Let $\mathcal{A} = \{\text{extra arm}, \text{extra hand}, \text{extra leg}, \text{extra foot}\}$. Given the HADM-G detection boxes $\mathcal{B}_t$ associated with the tracked person and the confidence $s_a(b)$ for each artifact category, we define

$$a_t = \max(\{s_a(b) : a \in \mathcal{A}, b \in \mathcal{B}_t\} \cup \{0\}), \qquad m_t^{\mathrm{a}} = \mathrm{clip}\left(\frac{a_t - \gamma_{\mathrm{lo}}}{\gamma_{\mathrm{hi}} - \gamma_{\mathrm{lo}}}, 0, 1\right). \tag{6}$$

The calibrated thresholds satisfy $0 \le \gamma_{\mathrm{lo}} < \gamma_{\mathrm{hi}}$. The union with $\{0\}$ assigns zero severity when no artifact box is predicted. Applicability annotations distinguish intended nonhuman transformations from anatomical defects.

**Rotation and motion quantities.** For each valid segment, let $\mathcal{J}$ denote the nonempty set of evaluated joints, $R_{j,t} \in \mathrm{SO}(3)$ the recovered local rotation of joint $j \in \mathcal{J}$, and $\Delta t > 0$ the sampling interval in seconds. Angular velocity, acceleration, and jerk are

$$\omega_{j,t} = \frac{\mathrm{Log}(R_{j,t+1} R_{j,t}^{\top})^{\vee}}{\Delta t}, \quad \alpha_{j,t} = \frac{\omega_{j,t} - \omega_{j,t-1}}{\Delta t}, \quad \jmath_{j,t} = \frac{\alpha_{j,t} - \alpha_{j,t-1}}{\Delta t}. \tag{7}$$

Here $\mathrm{Log}(\cdot)^{\vee}$ converts a relative rotation into a rotation vector under a consistent local coordinate convention for each joint. Local jerk energy is $E_{j,t} = |W_t|^{-1} \sum_{u \in W_t} \|\jmath_{j,u}\|^2$, where $W_t$ is a nonempty local window containing only indices with valid jerk estimates within the segment.

We adapt HumanScore's acceleration and jerk criteria (Fang et al., 2026) to the recovered joint rotations. Let $b_j^{\alpha}$ and $b_j^{E}$ be positive reference scales calibrated on coherent human motion in the annotated set. With the exceedance ramp $\rho(x) = \mathrm{clip}((x-1)/\kappa, 0, 1)$ and ramp width $\kappa > 0$, smoothness severity is

$$m_t^{s} = \frac{1}{2|\mathcal{J}|} \sum_{j \in \mathcal{J}} \left[ \rho\left(\frac{\|\alpha_{j,t}\|}{b_j^{\alpha}}\right) + \rho\left(\frac{E_{j,t}}{b_j^{E}}\right) \right]. \tag{8}$$

**Body and head heading.** We recover the forward directions of the body and head in world coordinates and project them onto the horizontal plane. Let $\psi_{h,t}$ denote the resulting heading for $h \in \{\text{body}, \text{head}\}$. Nearly vertical forward directions are excluded because their horizontal headings are unstable. Let $\mathcal{H}_t \subseteq \{\text{body}, \text{head}\}$ index the headings that are valid at both $t$ and $t+1$. At indices with $\mathcal{H}_t \neq \varnothing$, orientation severity is

$$m_t^o = \max_{h \in \mathcal{H}_t} \rho\left(\frac{|\operatorname{wrap}(\psi_{h,t+1} - \psi_{h,t})|}{b_h^\psi \Delta t}\right), \tag{9}$$

where wrap returns the angle in $(-\pi, \pi]$ and the positive reference scale $b_h^\psi$ is calibrated on valid turning motion.

**Aggregation.** For one reward component, let $m_t \in [0,1]$ be its severity at the nonempty set of eligible frame indices $T_S$ of segment $S$, omitting the component index for now. Define $b_t = \mathbf{1}[m_t > \delta]$ to flag an error. Let $L_{\max}$ be the longest run of flagged consecutive video frames, or zero if none are flagged. We compute error frequency, severity conditional on an error, and persistence as

$$f_S = \frac{\sum_{t \in T_S} b_t}{|T_S|}, \qquad s_S = \frac{\sum_{t \in T_S} b_t m_t}{\max(1, \sum_{t \in T_S} b_t)}, \qquad p_S = \frac{L_{\max}}{|T_S|}. \tag{10}$$

We set $D(S) = \nu_f f_S + \nu_s s_S + \nu_p p_S$, with nonnegative weights summing to one. For component $\ell \in \{\text{a}, s, o\}$, let $\mathcal{S}_\ell$ denote its eligible segments and $T_{S,\ell}$ its eligible frame indices within segment $S$. Applying the preceding aggregation to severity $m_t^\ell$ gives $D_\ell(S)$. For a component with at least one eligible frame, the score is

$$R_\ell = 1 - \frac{\sum_{S \in \mathcal{S}_\ell} |T_{S,\ell}| D_\ell(S)}{\sum_{S \in \mathcal{S}_\ell} |T_{S,\ell}|}. \tag{11}$$

Thus, $R_{\text{a}}$, $R_s$, and $R_o$ lie in $[0,1]$. The RL training subset is restricted to samples containing valid human-visible intervals for reward computation. We use $(\nu_f, \nu_s, \nu_p) = (0.5, 0.3, 0.2)$ to aggregate frequency, severity, and persistence.

### B.1 DiffusionNFT with Supervised Regularization

The frozen sampling policy $u_{\text{old}}$ generates $G = 16$ videos for each single-person input $x = (I, c, r)$ drawn from the subset of challenging effects. Let $R_i$ denote the human reward for video $i \in \{1, \ldots, G\}$ and $\overline{R}_x$ the group mean. The core DiffusionNFT objective (Zheng et al., 2026) uses soft positive weights

$$p_i = \frac{1}{2} + \frac{1}{2} \operatorname{clip}\left(\frac{R_i - \overline{R}_x}{\sigma_R + \varepsilon_{\text{num}}}, -1, 1\right), \tag{12}$$

where $\sigma_R$ is the standard deviation across the rollout batch and $\varepsilon_{\text{num}}$ stabilizes the denominator. Thus, rewards are centered within each group and scaled across the batch. At a training timestep $t \in [0,1]$, we perturb the generated clean latent as $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\xi$, with $\xi \sim \mathcal{N}(0, \mathbf{I})$ and velocity target $y = \xi - \mathbf{z}_0$. Let $u_\theta$ denote the current velocity predictor. At the same $(\mathbf{z}_t, t, x)$, define

$$u_\theta^+ = (1-\beta)u_{\text{old}} + \beta u_\theta, \qquad u_\theta^- = (1+\beta)u_{\text{old}} - \beta u_\theta. \tag{13}$$

The objective is

$$\mathcal{L}_{\text{NFT}} = \mathbb{E}\left[p_i \|u_\theta^+ - y\|_{\text{gen}}^2 + (1 - p_i)\|u_\theta^- - y\|_{\text{gen}}^2\right]. \tag{14}$$

The expectation averages over input conditions, generated rollouts, training timesteps, and Gaussian perturbations. Here, $\|e\|_{\text{gen}}^2$ denotes the mean squared error over generated latent coordinates. Gradients are stopped through the rewards and $u_{\text{old}}$. After each round, we update the sampling policy using an exponential moving average of the video adapter.

**Supervised replay.** Each update uses an independent supervised batch from the full 100K single-person effect corpus. The supervised term in Eq. 3 applies the original flow-matching loss with the same reference preparation and normalization over generated latent coordinates. This data regularization (Ye et al., 2025) preserves effect appearance and dynamics during reward optimization. The ablation removes only $\mathcal{L}_{\text{SFT}}$, keeping the reward, rollout budget, and initialization unchanged.

## C VLM-GUIDED TEST-TIME OPTIMIZATION

Algorithm 1 optimizes the initial noise while keeping the distilled generator, LoRA adapters, and VLM fixed. For multi-person inputs, we disable the identity branch and condition generation on image $I$ and prompt $c$. Prediction depth $k \in \{1, \dots, S\}$ sets the number of denoising evaluations per assessment, $U$ limits the noise updates, and $S$ sets the final sampling steps. We use $k = 1$, $U = 4$, and $S = 4$ by default.

**Clean-latent estimation.** Let $1 = t_0 > \cdots > t_S = 0$ denote the distilled sampler's schedule. Starting from $\mathbf{z}_{t_0} = \mathbf{z}$, we take $k - 1$ scheduled denoising steps and use the $k$th velocity prediction to estimate the clean latent:

$$\hat{\mathbf{z}}_0^{(k)}(\mathbf{z}; I, c) = \mathbf{z}_{t_{k-1}} - t_{k-1} u_{\theta,\phi}(\mathbf{z}_{t_{k-1}}, t_{k-1}, I, c), \tag{15}$$

where $u_{\theta,\phi}$ is the fixed generator with its distilled adapter. At $k = 1$, this reduces to $\hat{\mathbf{z}}_0^{(1)}(\mathbf{z}) = \mathbf{z} - u_{\theta,\phi}(\mathbf{z}, 1, I, c)$, following the first-step prediction of Cheng et al. (2026). For larger $k$, we apply the same estimate after additional denoising steps. We restore the fixed input-frame latent and decode the estimate with the lightweight surrogate decoder from LightX2V (LightX2V Contributors, 2025), producing the intermediate video $\tilde{\mathbf{v}}(\mathbf{z})$ in Eq. 5. The VLM assesses $K$ sampled frames. All $k$ denoising evaluations remain in the computation graph so that VQA gradients reach the initial noise.

**Initial noise updates.** Norm-preserving spherical updates adjust the noise direction while keeping its energy constant. The noise variable and its norm include only generated latent coordinates, excluding the first-frame condition. We optimize the VQA loss in Eq. 4 without an additional latent regularization term.

**Algorithm 1** VLM-guided test-time optimization

**Require:** image $I$, prompt $c$, frozen generator and VLM $P_\psi$, decoders $\mathcal{D}_{\text{lite}}$ and $\mathcal{D}$, prediction depth $k$, update budget $U$, threshold $\tau_L$, step size $\eta$, final sampling depth $S$

$\mathcal{Q} \leftarrow \text{AssessRequirements}(P_\psi, I, c)$
Sample initial noise $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and set $\rho \leftarrow \|\mathbf{z}\|_2$
**for** $n = 0, \dots, U - 1$ **do**
  Estimate $\hat{\mathbf{z}}_0^{(k)}(\mathbf{z}; I, c)$ using Eq. 15
  $\tilde{\mathbf{v}} \leftarrow \mathcal{D}_{\text{lite}}(\hat{\mathbf{z}}_0^{(k)})$
  $L \leftarrow \mathcal{L}_{\text{VQA}}(\tilde{\mathbf{v}})$ using questions $\mathcal{Q}$
  **if** $L \le \tau_L$ **then**
    **break**
  **end if**
  $\mathbf{z} \leftarrow \text{SphericalStep}(\mathbf{z}, \nabla_{\mathbf{z}} L, \eta, \rho)$
**end for**
**return** $\mathcal{D}(G_{\theta,\phi}^{(S)}(I, c; \mathbf{z}))$

**Optimization and final generation.** The VLM chooses one to five equally weighted assessment questions for each input. We fix the questions and temporal frame indices throughout optimization. After differentiable resizing and normalization, the VLM evaluates the target token sequence for "Yes" under teacher forcing. Gradients propagate through the VLM, lightweight decoder, and denoiser. Optimization stops when the mean VQA loss reaches $\tau_L = 0.22$ or the update budget is exhausted. Finally, $G_{\theta,\phi}^{(S)}$ produces the clean video latent from the refined noise in $S = 4$ sampling steps, and the original VAE decoder produces the output video.

### C.1 Visual Assessment Prompts

We generate the assessment questions with the following instruction:

**Assessment-question generation prompt**

Examine the input image and video description. Identify the people by their initial positions and visible attributes. Write one to five focused yes/no questions that assess the requested effects, target people, actions, reactions, and interactions. Choose the number of questions needed to cover the instruction without repetition. Every question must have Yes as its desired answer. Include event order when the description specifies it. For a selective effect, ask whether only the designated person receives that effect. Ask only about requirements supported by the description, and do not add unspecified appearance or scene-preservation requirements. Return only the questions.

### C.2 Computational Cost

We measure end-to-end wall time on the same device and at the same numerical precision. TTO timing covers question preparation, intermediate predictions, lightweight decoding, VLM assessment, backward passes, and final four-step generation. Best-of-$N$ timing covers the same question preparation, generation of $N$ complete videos, and their VQA assessment. The independent evaluator used to report joint success is excluded from both generation times.

Generation cost is each method's wall time divided by the time for one four-step generation with the standard decoder, expressing inference overhead relative to a single sample. The shared baseline has cost 1 and requires no VLM assessment. Figure 4 uses $k = 1$ and $S = 4$, with TTO budgets of zero, one, two, and four updates and best-of-$N$ with one, four, and six candidates. Timing includes only the assessments and updates performed before stopping.

## D Data and Implementation Details

**Data preparation.** We select single-person clips from Open-VFX (Liu et al., 2025b), VAP-Data (Bian et al., 2026), and EffectData (Yang et al., 2026), retaining clips that follow one human subject through the effect and removing near-duplicates across training and evaluation. The resulting corpus contains 100,000 unique clips with effects such as environmental changes, subject transformations, stylization, camera effects, and composite effects. Given the first frame and video, Qwen3.5-122B-A10B (Qwen Team, 2026) describes the subject, effect progression, actions, camera movement, and appearance or scene elements that remain unchanged.

### D.1 Training and Inference Settings

Table 3 summarizes the training and inference settings.

We train with AdamW (Loshchilov & Hutter, 2019). SFT updates both LoRA adapters, while RL updates the video adapter and freezes the identity adapter. Training preserves the original aspect ratio under a 720p pixel budget and supports variable clip lengths up to 241 frames. RL uses selected single-person effects with large or complex movements, pairing each rollout minibatch with an equally sized supervised minibatch from the broader corpus. Reward ablations share the calibration, inputs, and sampling settings. Evaluation preserves the input aspect ratio under a $704 \times 1280$ pixel budget, except for baselines that require a fixed native resolution.

**Alternate-view face reference.** For single-person inference, we prompt Qwen-Image-Edit (Wu et al., 2025) to synthesize a complementary facial viewpoint while requesting preservation of identity, age, hairstyle, and expression. We align and crop the edited face as during training, then reuse this reference across samples. It conditions the identity branch, while the original image remains the video's first frame. We disable the identity branch for multi-person inputs.

**Distillation and visual assessment.** The RL-trained model is distilled into a four-step generator following DMD2 (Yin et al., 2024). Qwen3-VL-4B (Bai et al., 2025) provides the VQA reward with the prediction and optimization procedure in Appendix C.

Table 3: Training and inference settings.

| Setting | Value |
|---|---|
| Backbone | Wan2.2-TI2V-5B |
| Video LoRA | Rank 256, Q/K/V/O and FFN |
| Identity LoRA | Rank 128, reference Q/K/V |
| Reference crop / tokens | $512 \times 512$ / 256 |
| SFT steps / learning rate | 4,000 / $5 \times 10^{-5}$ |
| RL steps / learning rate | 50 / $1 \times 10^{-5}$ |
| Batch size / training hardware | 32 / 32 NVIDIA H20 GPUs |
| Optimizer / weight decay | AdamW / 0.01 |
| Caption / reference dropout | 0.1 / 0.2 |
| Gradient clipping | 1.0 |
| Training resolution budget | 720p |
| Maximum training clip length | 241 frames |
| Rollout group / NFT $\beta$ / EMA decay | 16 / 0.1 / 0.9 |
| SFT weight / rollout-to-data batch ratio | 0.1 / 1 : 1 |
| Reward weights $(\lambda_{\mathrm{a}}, \lambda_{\mathrm{m}})$ | $(0.5, 0.5)$ |
| Severity ramp width / flag threshold | 0.5 / 0.05 |
| Aggregation $(\nu_f, \nu_s, \nu_p)$ | $(0.5, 0.3, 0.2)$ |
| Initial noise update | Norm-preserving spherical step |
| Test-time learning rate / maximum updates $U$ | 0.01 / 4 |
| Update budgets in ablation | 0, 1, 2, 4 |
| Prediction depth $k$ / sampled frames $K$ | 1 / 40 |
| Stopping loss averaged over questions | 0.22 |
| VQA questions / weighting | 1–5 / equal |
| Test-time VQA model | Qwen3-VL-4B |
| Final sampling steps $S$ | 4 |
| HumanFX output length / frame rate | 121 frames / 24 fps |

# E Evaluation Details

## E.1 Data and Comparisons

We adapt the Open-VFX evaluation protocol used by EffectMaker (Yang et al., 2026), with 10 held-out human images for each of 14 effects, yielding 140 image–prompt pairs for the main comparison and ablations. Effect quality is assessed over the complete sequence, while human preservation is assessed where the instruction requires the subject to retain human form.

The multi-person set contains 60 image–prompt pairs covering selective and joint effects, each specifying the recipients and associated actions, reactions, or interactions. We generate three outputs per input, giving 420 single-person videos and 180 multi-person videos per method. Open models use fixed seeds, while the commercial baseline receives three independent requests. Evaluation identities, source images and videos, near-duplicates, and derived variants are excluded from training and reward calibration.

**Comparison settings.** All methods receive the same initial image and effect description. VFX Creator uses the corresponding effect adapter and any required spatial mask annotated on the image. Omni-Effects uses the released prompt-guided `CogVideoX1.5-5B-I2V-OmniVFX` checkpoint. MiniMax-H3 uses its hosted first-frame image-to-video mode (MiniMax, 2026a) at 768P for five seconds. HumanFX produces 121 frames at 24 fps, and baselines retain native output settings where necessary. VLM evaluation samples frames uniformly over the full duration. VBench uses its released preprocessing, including frame-rate-aware subsampling for dynamic degree. Preprocessing and human motion reconstruction retain native timestamps without synthesizing intermediate frames.

**Test-time comparisons.** All multi-person variants share the trained model, four-step sampler, and VQA questions and reward, with the identity branch disabled. Best-of-$N$ selects the complete candidate with the lowest VQA loss, while TTO uses its gradient to refine the initial noise. An independent evaluator assesses the final outputs. Runtime measurement follows Appendix C.2.

### E.2 METRICS

**VideoReward.** Following EffectMaker (Yang et al., 2026), we evaluate visual quality (VQ) and motion quality (MQ) with VideoReward (Liu et al., 2025a). VQ covers clarity, aesthetics, and single-frame quality. MQ combines temporal stability, motion naturalness, and dynamic extent. We use the released checkpoint and official normalization, standardizing each dimension with the corresponding mean and standard deviation. The resulting scores are unbounded and may be negative. The evaluator samples the complete video at a target density of 2 fps, preserving aspect ratio with at most $200{,}704$ pixels per frame. VQ and MQ assess appearance and movement independently of prompt-specific effect requirements. We interpret MQ alongside dynamic degree and alignment because natural motion alone does not establish whether a nonphysical effect is correct.

**Motion smoothness and dynamic degree.** Following Video-As-Prompt (Bian et al., 2026), we report VBench motion smoothness and dynamic degree (Huang et al., 2024). Smoothness compares withheld intermediate frames with predictions from a frame-interpolation model. Dynamic degree is the percentage of clips classified as dynamic by the optical-flow criterion, measuring motion prevalence rather than average amplitude. Both are needed because a nearly static video can be smooth while leaving a dynamic effect incomplete (Liao et al., 2024).

**Effect alignment.** GPT-6 Astra (OpenAI, 2026) scores a uniformly sampled sequence of frames from each generated video against the effect instruction using the rules in Appendix F, adapted from Video-As-Prompt (Bian et al., 2026). The score ranges from $0$ to $100$ and covers effect execution, subject assignment, and event progression. All methods share the frame sampling and scoring instructions. Scores are averaged over the three outputs per input and then over inputs.

**Face similarity.** Following Stand-In (Xue et al., 2026), FaceSim measures cosine similarity between CurricularFace (Huang et al., 2020) embeddings of the input face and detected faces in $32$ uniformly sampled video frames. We use aligned face crops and the original input image as the reference. The synthesized alternate view is used only for generation. We average cosine similarities over valid face detections and assign zero if none remain. We then average over the three outputs per input and across inputs, retaining the cosine-similarity scale.

**Reward diagnostics.** The ablation study reports $R_{\mathrm{a}}$ and $R_{\mathrm{m}}$ from Eqs. 2 and 11, scaled by $100$. Following HumanScore (Fang et al., 2026), motion analysis focuses on recoverable human movement. We compute these rewards within intervals where the subject should retain human form. Missing motion fits within eligible intervals receive maximal severity, and fitting coverage is recorded. Scores are averaged over outputs per input and then over inputs.

**Multi-person control.** Assignment succeeds when every designated person receives the requested effect and selective effects do not spread to others. Behavior succeeds when the requested actions, reactions, and interactions occur in the specified order. Joint success requires both. GPT-6 Astra (OpenAI, 2026) applies these binary criteria to all $180$ outputs, and each metric reports the percentage of successful outputs. Intended changes and disappearance are evaluated as part of the requested event. The evaluator receives the complete instruction without the test-time questions.

### E.3 HUMAN PREFERENCE STUDY

Following Video-As-Prompt (Bian et al., 2026), we report each method's share of user selections. We recruit 20 participants, six with experience in design or film and video production. Each evaluates the same $72$ image–prompt pairs, yielding $1{,}440$ selections. Each trial presents the input image, description, and one preselected video from each of the five methods. Method names are hidden and display order is randomized. Participants can replay the videos before choosing the result that best combines effect fidelity, identity preservation, coherent human appearance and motion, scene continuity, and visual quality. Each selection gives one vote to one method. Preference is its share of all $1{,}440$ votes, rounded to one decimal place, so the unrounded shares sum to $100\%$.

# F VLM EVALUATION PROTOCOL

GPT-6 Astra (OpenAI, 2026) (`gpt-6-astra`) receives the effect instruction together with uniformly sampled frames from the generated video and their timestamps. We adapt Video-As-Prompt's semantic assessment (Bian et al., 2026) to score effect execution, subject assignment, and event progression, worth up to 60, 20, and 20 points, respectively. Align. is their sum, capped at 20 if the principal effect is absent or applied to the wrong subject. The following prompt defines the evaluation rules.

**Effect-alignment evaluation prompt**

Evaluate how well the video generates the visual effects described in the instruction. Use the opening frame to identify the people and scene. Judge the requested events from the sampled frames and their timestamps.

Give three integer scores: effect execution from 0 to 60, subject assignment from 0 to 20, and event progression from 0 to 20. Assess the requested effect appearance and extent, the people or regions to which it applies, and the development of the event over time. Evaluate any requested reactions and interactions as part of event progression. Do not require an onset, reversal, or return to the initial state unless the instruction calls for it.

Allow changes specified by the effect, such as stylization, disappearance, and nonhuman transformations. Assess preservation where the instruction requires it. People who retain their appearance may still move or react. Distinguish an attractive video from one that correctly executes the instruction.

If the principal effect is absent or applied to the wrong subject, cap the total score at 20. Return the three component scores, their sum after applying the cap, and one short explanation tied to visible events.