FROM COMPOSITES TO CONTROL: GEOMETRY-GROUNDED SUPERVISION FOR VISUAL EFFECTS GENERATION

Anonymous authors
Paper under double-blind review

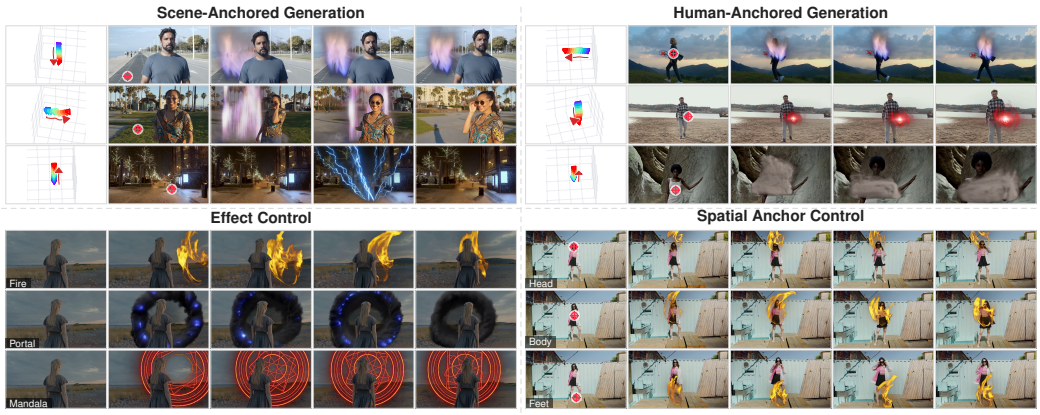

Figure 1: Controllable VFX generation with VFXDirector. Given a first-frame image, an effect prompt, a camera trajectory, and a spatial anchor with its relation, VFXDirector generates videos following the specified effect and geometric controls. Top: representative generation results. Bottom: effect and geometry control under different effect prompts and human-anchor relations.

ABSTRACT

Recent visual effects (VFX) generation methods primarily rely on large video generative models to synthesize both effects and their interactions with scenes. While this paradigm enables flexible creation, its reliance on uncertain generative priors makes effect fidelity and geometric controllability difficult to ensure, especially under camera or subject motion. We introduce effect-faithful geometry-grounded VFX generation, a formulation that decouples reliable VFX supervision from generative uncertainty. Instead of using a video generator to synthesize VFX videos, we composite curated VFX assets onto real-world footage using recovered camera motion, scene geometry, and human motion. This preserves natural video dynamics and asset-defined effect appearance while providing explicit geometric supervision. Based on this principle, we build **VFXShot**, a dataset of 10,298 videos with effect semantics, camera trajectories, and spatial anchor annotations. We further propose **VFXDirector**, a unified video generation framework that conditions on effect semantics, camera trajectories, and spatial attachment. Experiments demonstrate that effect-faithful geometry-grounded supervision improves controllable VFX generation, enabling VFXDirector to achieve improved video quality, camera consistency, and spatial attachment. These results suggest that reliable compositional supervision provides an effective alternative to purely generative data construction for learning controllable visual effects. Code is available at Anonymous Repository.

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

1 INTRODUCTION

Recent video generation models have made substantial progress in visual fidelity, temporal coherence, and text-conditioned content creation (Guo et al., 2024b; Chen et al., 2024; Yang et al., 2025; Wan et al., 2025). These advances have also enabled visual effects (VFX) generation (Guo et al., 2024a), where effects such as fire, lightning, smoke, and explosions are synthesized directly into videos. Existing methods extend pretrained video generation models with textual descriptions, spatial masks, temporal conditions, or reference effects (Liu et al., 2025; Mao et al., 2026; Yang et al., 2026). However, the models must simultaneously determine what the effect looks like and how it interacts with the scene. Both are inferred from the generative prior under the supplied conditions. This creates two coupled challenges: *effect fidelity*, which concerns preserving the intended effect appearance and dynamics, and *geometric controllability*, which concerns following the specified camera motion and spatial relationship. For example, fire may need to stay fixed on the ground while the camera moves around it, whereas electrical energy may need to follow a moving hand.

Existing VFX data construction pipelines provide limited supervision for these coupled requirements. Previous works such as Open-VFX (Liu et al., 2025), Omni-VFX (Mao et al., 2026), and EffectData (Yang et al., 2026) obtain target videos from generative systems or image-editing and video-generation pipelines. The effect appearance and internal dynamics are therefore model-generated rather than explicitly specified by VFX assets. These datasets provide effect categories, spatial regions, temporal annotations, or reference appearances, but not joint camera trajectory and spatial anchoring supervision for scene- or human-anchored effects. Learning how a requested effect should evolve within a moving shot thus calls for data that specifies both its visual content and its spatial relationships.

Our key insight is to separate reliable VFX supervision from generative uncertainty. Rather than asking another video generator to synthesize both the scene and the effect, we deterministically composite curated VFX assets onto real-world footage using recovered geometry. The original video provides natural camera, scene, and human dynamics, while the asset defines the effect appearance and internal animation. Geometry connects the two. Our method estimates camera motion, scene depth, and human pose to determine placement, scale, and visibility. *Scene-Anchored* effects are tied to reconstructed scene locations and reprojected as the camera moves, whereas *Human-Anchored* effects follow semantic body locations. Foreground occlusion is handled during scene-anchored compositing. Using this method, we construct **VFXShot**, containing 10,298 VFX videos from 542 real source videos and 19 effect types, with aligned effect, camera, and spatial annotations. This establishes a path from composites to control, where compositing provides effect-faithful, geometry-grounded supervision for learning controllable VFX generation.

We further introduce **VFXDirector**, a unified video generation framework that learns these relationships from VFXShot. Given a first-frame image, an effect description, a target camera trajectory, and a spatial anchoring condition, VFXDirector generates the entire VFX video without requiring an explicit VFX asset at inference time. The anchoring condition combines a first-frame 2D point and an anchoring relation with no future anchor trajectory required. Camera motion and spatial anchoring together determine how the effect should move in image space. VFXDirector therefore encodes camera and anchor conditions and jointly integrates them throughout the video diffusion transformer through adaptive feature modulation. Text conditioning specifies the requested effect and anchoring relation, supporting effect selection. Extensive experiments demonstrate that fine-tuning on VFXShot improves VFX video generation over the pretrained backbone. VFXDirector further reduces FVD and improves camera trajectory and spatial anchor adherence. These results validate the effectiveness of our approach and demonstrate that explicit effect construction and geometric supervision can support unified controllable VFX generation.

Our contributions are summarized as follows:

- We introduce a VFX data synthesis method that deterministically composites curated effects onto real footage using recovered geometry, producing effect-faithful and geometry-grounded supervision.
- We construct VFXShot, a dataset of 10,298 videos covering 19 effect types, with paired camera trajectories and scene- or human-anchored spatial annotations.

- We instantiate VFXDirector to validate the effectiveness of VFXShot. It conditions on effect semantics, camera trajectories, and spatial anchoring through text conditioning and joint adaptive feature modulation.

2 Related Work

General Video Generation. Diffusion-based video generation has evolved from U-Net and latent-diffusion architectures with temporal modeling and motion adaptation (Ho et al., 2022; Singer et al., 2023; Blattmann et al., 2023; Guo et al., 2024b; Chen et al., 2024; Xing et al., 2024) to scalable Diffusion Transformer (DiT) and flow-based models (Peebles & Xie, 2023; Ma et al., 2025; Yang et al., 2025; Jin et al., 2025; Lin et al., 2025). Recent large-scale models further strengthen general video generation priors (Kong et al., 2024; Wan et al., 2025; HaCohen et al., 2026; Ma et al., 2026). Building on Wan, we adapt such priors to geometry-grounded VFX generation with explicit camera and spatial control.

Visual Effects Generation. Recent VFX generation methods study effect adaptation, spatial and temporal control, multi-effect composition, and training-free customization (Liu et al., 2025; Mao et al., 2026; Zhao & Shou, 2026), while reference-conditioned methods transfer semantic or dynamic effect information from example videos (Bian et al., 2026; Li et al., 2026; Yang et al., 2026). These methods often rely on image-editing and video-generation pipelines to construct VFX videos (Liu et al., 2025; Mao et al., 2026; Yang et al., 2026). Unlike these methods, we composite curated VFX assets onto real-world footage using recovered geometry, and the model jointly controls an explicit camera trajectory and persistent scene- or human-attachment from a sparse first-frame anchor and its relation.

Controllable Video Generation. Controllable video generation uses structural, trajectory, and camera conditions to regulate spatial and temporal dynamics (Wang et al., 2023; Jiang et al., 2025; Yin et al., 2023; Shi et al., 2024; Wu et al., 2024; Zhang et al., 2025b; Geng et al., 2025; Fu et al., 2025). Recent methods further support object–camera motion control or explicit camera conditioning through camera parameters and geometric representations (Wang et al., 2024b; Yang et al., 2024; Zheng et al., 2024; Shin et al., 2026; He et al., 2025a; Feng et al., 2025; He et al., 2025b; Bai et al., 2025a; Zhang et al., 2025a). Our method instead uses only a first-frame anchor and its relation for spatial attachment, together with an explicit camera trajectory, without requiring future effect trajectories.

3 VFX Data Synthesis

3.1 Overview

Learning controllable VFX generation requires supervision for both effect fidelity and geometric controllability. Geometry-aware visual modeling has studied depth and viewpoint information for reconstructing scene structures (Wang et al., 2024a; Zhao et al., 2024). However, existing VFX datasets obtain their target VFX videos from generative pipelines with image-editing and video-generation models (Liu et al., 2025; Mao et al., 2026; Yang et al., 2026). In these pipelines, effect appearance is itself model-generated, and the resulting data do not provide explicit geometric information such as camera trajectory and spatial anchoring supervision. We address both aspects during data construction, where curated VFX assets specify the effect appearance and internal animation, while recovered video geometry determines their spatial placement. This produces effect-faithful, geometry-grounded supervision from explicit effect content and placement rules.

Our synthesis pipeline combines geometric reconstruction, semantic analysis, and deterministic VFX compositing, as illustrated in Fig. 2. Instead of generating the entire target video, we preserve the original camera motion, scene dynamics, and human motion, and composite the selected VFX asset onto the source footage. Spatial anchoring specifies both the location of an effect and the reference that governs its motion over time. We consider two forms: **Scene-Anchored** effects are tied to a location in the reconstructed scene, whereas **Human-Anchored** effects follow a semantic body location (e.g., head, body, hands, or feet). The method prepares the inputs, specifies the effect, computes its placement through either anchoring branch, and combines the resulting animation and

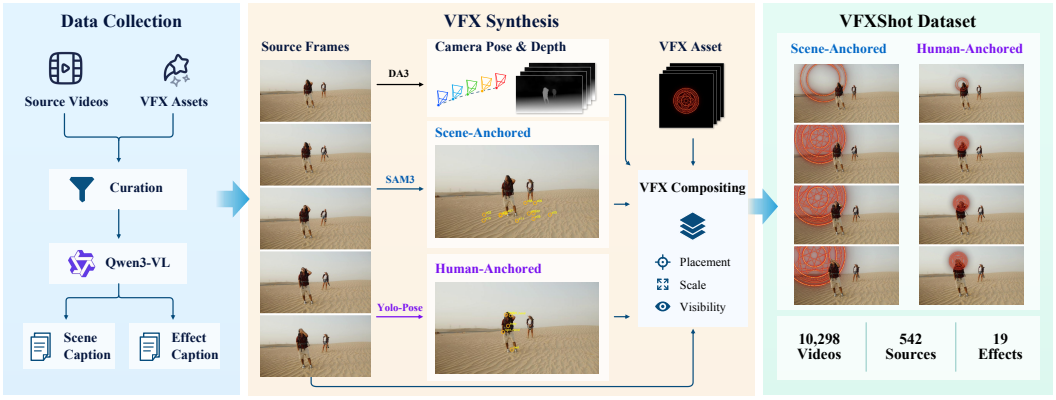

Figure 2: **Overview of the VFXShot dataset construction pipeline.** We collect real-world source videos and VFX assets, and then synthesize VFX videos with aligned effect, camera, and spatial information. The pipeline consists of source collection and filtering, geometric and semantic processing, spatial anchor generation, VFX synthesis, and quality control.

geometry through a shared compositing procedure. We first describe this method and then introduce the resulting VFXShot dataset.

3.2 METHOD

Inputs. To enhance the diversity of our dataset, we collect real-world source videos and animated VFX assets as inputs. Source clips are manually screened for visual quality and suitable effect locations. We retain footage with clear scene structure and regions that support consistent placement. Human-anchored synthesis additionally requires visible and trackable body regions. We use Depth Anything 3 (DA3) (Lin et al., 2026) to estimate depth, confidence, camera intrinsics, and camera extrinsics, which provide the geometric basis for placement and perspective-dependent scaling. Qwen3-VL (Bai et al., 2025b) generates separate descriptions of the source video and VFX asset, providing scene and effect semantics for the synthesized sample.

Effect Specification. Each synthesized sample uses a selected VFX asset as the explicit source of its effect appearance and internal animation. The asset contains RGB frames and an alpha channel for compositing. We associate each asset with placement settings for the two anchoring modes, including an alignment point within the asset and a valid scale range. The alignment point specifies which part of the asset should coincide with the scene or human anchor, while the scale range determines its intended screen coverage. These settings allow the same effect to be placed in different spatial contexts without changing its identity. Effect-faithful supervision thus comes from reusing the selected high-quality VFX assets rather than asking a video generator to recreate VFX appearance for each source shot.

Scene-Anchored. For scene-anchored synthesis, we construct a pool of candidate 2D anchor points through three stages of filtering: semantic regions, local geometry, and cross-frame consistency. We use SAM3 (Carion et al., 2026) to segment the scene with support regions S_{support} (e.g., road, ground, open area) and avoid regions S_{avoid} (e.g., person, animal, sky), and identify candidate anchor points in semantically meaningful regions (i.e., $S_{\text{support}} \setminus S_{\text{avoid}}$). The candidates are then filtered using depth confidence and local depth continuity. Using the estimated depth and camera parameters, the candidate points are lifted to 3D scene locations and checked for multi-frame visibility and projected effect size. This makes placement selection depend on both the scene geometry and the chosen effect's scale. The selected scene point is reprojected into subsequent frames using the recovered camera trajectory, keeping the VFX tied to the same scene location as the viewpoint changes. The resulting projections provide the frame-wise positions used for compositing.

Human-Anchored. We use YOLO-Pose (Jocher et al., 2026) to track the target person and construct 2D anchor sequences for the head, body, left hand, right hand, and feet. The selected semantic location defines the anchoring relation, while its tracked coordinates determine the effect's position in each frame. These coordinates follow the corresponding body region, so their image-space motion naturally combines human motion with camera motion. Short tracking gaps are repaired when reliable interpolation is possible, while discontinuous or invalid tracks are rejected during construction. We estimate a shared body depth from the torso region and smooth it over time to stabilize perspective-dependent effect scale. The original 2D motion of each semantic location is retained. The accepted trajectory and depth sequence are then passed to the shared compositing procedure.

Compositing. For both anchoring modes, we align the asset's specified alignment point with the projected or tracked position and adjust its scale using the placement settings and estimated depth. For scene-anchored effects, we additionally combine SAM3 foreground masks with depth information to identify scene objects in front of the effect. Occluded portions of the VFX are suppressed during alpha compositing, preserving the foreground ordering of the source footage. A gradual temporal onset controls how the effect becomes visible, while its internal animation comes from the selected asset. With the source footage, asset, estimated geometry, and placement configuration fixed, this procedure deterministically produces the composite video. The camera parameters, anchor positions, and anchoring relation are recorded from the same construction process, establishing a direct correspondence between the target video and its spatial supervision.

This construction makes the resulting videos effect-faithful as the selected asset, rather than a generative model, specifies the effect's appearance and internal animation. The geometry-aware compositing process changes only how the asset is placed, scaled, and occluded in the source footage. Each target video therefore preserves the specified effect identity while exposing it to diverse real-world scenes and motions.

3.3 Dataset

Using the proposed method, we construct VFXShot, containing 10,298 VFX videos synthesized from 542 real-world source videos. The source footage is collected from Pexels and covers indoor and outdoor scenes as well as human activities. We use 19 animated VFX assets spanning fire, smoke, water, lightning, and magical effects, with 19 effect variants generated for each source video. The dataset includes 5,719 Scene-Anchored videos from 301 source videos and 4,579 Human-Anchored videos from 241 source videos. All final videos are stored at 720p and 24 fps. By pairing different effects with the same recorded scene and motion, VFXShot provides structured examples of effect variation within real dynamic footage, together with the camera and anchoring information produced by the synthesis method. For more details, please refer to Appendix A.

4 Model

4.1 Overview

Given the structured information provided by VFXShot, we develop VFXDirector, a unified framework for controllable VFX generation (See Fig. 3). The task is formulated as follows: given a first-frame image I_0, an effect description e, a camera trajectory $C_{0:T}$, and a spatial anchoring condition $\mathcal{A}_0$, VFXDirector generates a video $Y_{0:T}$ that follows the requested effect, viewpoint evolution, and spatial placement. We define $\mathcal{A}_0 = (\mathbf{p}_0, r)$, where $\mathbf{p}_0$ is a 2D anchor point in the first frame and r denotes its anchoring relation. Only the first-frame anchor point is required at inference time, and no future anchor trajectory is provided. We build VFXDirector on Wan2.2-TI2V-5B (Wan et al., 2025) as the backbone. We describe the control condition encoding and the joint adaptive fusion module as follows.

4.2 Control Encoding

Camera Encoding. Camera motion determines how the entire shot evolves over time and directly affects the apparent motion of both scene content and VFX. We provide the target camera trajectory as a sequence of camera extrinsics following previous work (Bai et al., 2025a). We convert the

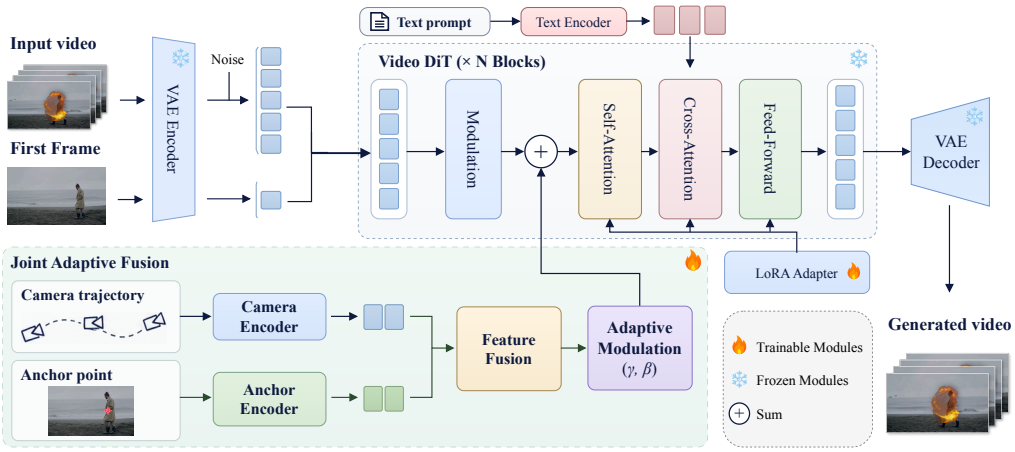

Figure 3: **Overview of VFXDirector.** VFXDirector generates videos jointly conditioned on a first-frame image, a camera trajectory, a first-frame 2D anchor point, and a text prompt specifying the effect and anchoring relation. Joint Adaptive Fusion combines encoded camera and anchor features into token-wise scale and shift parameters that modulate each DiT block before self-attention.

camera trajectory so that each camera pose is relative to the first frame. For a camera trajectory with f frames, we represent each camera pose at frame t by the extrinsic matrix $E_t \in \mathbb{R}^{3\times4}$. Then the trajectory is temporally aligned with the latent video frames, and each relative 3×4 extrinsic matrix is flattened into a 12-D pose representation P. Instead of converting the complete trajectory into a single global embedding, we retain its temporal structure and inject it into every DiT block. For block l, a camera encoder E_l^{c} projects the pose sequence to the hidden dimension and broadcasts each temporal representation over the corresponding spatial tokens, producing camera features $C_l = E_l^{\mathrm{c}}(P)$ aligned with the visual token sequence. This block-wise encoding allows camera information to directly condition visual representations throughout the denoising network.

Spatial Anchor Encoding. The spatial anchor encoding captures the initial location and relationship of the VFX within the scene. The 2D anchor point $\mathbf{p}_0$ specifies *where* the effect is initialized, while the relation r specifies *what* the effect is anchored to. This interface supports both scene-anchored and human-anchored generation within a single model. A 2D coordinate alone indicates where an effect should appear, but provides little information about the local visual content at that position. We therefore encode the anchor using both its spatial location and the appearance of the first frame. Given the normalized point $\mathbf{p}_0$, we construct a Gaussian map centered at the anchor, with non-zero support only at the first latent frame. We then bilinearly sample the clean first-frame latent at $\mathbf{p}_0$ to obtain a local appearance vector and distribute this feature according to the Gaussian map. The spatial Gaussian and appearance feature map are concatenated and processed by a lightweight 3D convolutional fusion module, yielding token-aligned anchor features F_a. A separate anchor encoder E_l^{a} in each DiT block maps F_a to the block hidden dimension, producing $A_l = E_l^{\mathrm{a}}(F_a)$.

The anchoring relation r is represented through the structured text prompt. Specifically, the text condition combines the video description, anchoring relation, and desired effect description in the form of $[T_{\text{video}}, T_{\text{relation}}, T_{\text{effect}}]$. The relation text is "Add this visual effect attached to the person's [keypoint]:" for human-anchored effects and "Add this visual effect fixed at the provided scene anchor:" for scene-anchored effects.

4.3 Joint Adaptive Fusion

Camera motion and spatial anchoring jointly determine how a VFX should evolve in image space. We therefore fuse the two conditions before self-attention and use the adaptive modulation inspired by previous work (Zhang et al., 2025b). At the l-th DiT block, let h_l denote the visual features after the original normalization and timestep modulation. The block-specific camera and anchor representations C_l and A_l are first projected into a compact control space and combined to form a

joint condition

$$z_l = \mathrm{SiLU}\left(D_l^c(C_l) + D_l^a(A_l)\right), \tag{1}$$

where D_l^c and D_l^a are lightweight camera and anchor projections. A joint projection U_l^{joint} then maps z_l to token-wise scale and shift parameters $[\gamma_l, \beta_l]$. We use these parameters to adapt the visual representation before self-attention:

$$\widetilde{h}_l = h_l + \mathrm{LN}(h_l) \odot \gamma_l + \beta_l. \tag{2}$$

The modulated features $\widetilde{h}_l$ are subsequently processed by the original self-attention, cross-attention, and feed-forward layers of the Wan block. Together, the structured text condition, camera trajectory encoding, spatial anchor encoding, and joint adaptive fusion form VFXDirector. This unified formulation enables different VFX, camera trajectories, and spatial anchors to be specified within the same generation model.

5 Experiments

5.1 Experimental Setup

Dataset and Baselines. The experiments are conducted on our proposed dataset VFXShot. To avoid data leakage, we ensure that the variants of the same source video are only present in one of the splits. We compare our method with state-of-the-art video generation models, as existing VFX methods use different control interfaces from our setting. The baselines include: (i) CogVideoX (Yang et al., 2025), (ii) Wan2.2 (Wan et al., 2025), (iii) LTX-2 (HaCohen et al., 2026), (iv) LingBot-Video (Ma et al., 2026), (v) we also finetune Wan2.2 on our dataset, denoted as Wan2.2-FT.

Evaluation Metrics. We evaluate model performance from three perspectives: (i) video quality, (ii) camera motion controllability, and (iii) spatial anchoring controllability. For video quality, we use Fréchet Video Distance (FVD) (Unterthiner et al., 2018) and VBench (Huang et al., 2024) metrics to evaluate imaging quality, subject consistency, and dynamic degree following previous works (Liu et al., 2025; Mao et al., 2026). For camera motion controllability, we use CamMC (Wang et al., 2024b), TransErr and RotErr (He et al., 2025a) to measure similarity between generated and ground-truth camera motions. For spatial anchoring controllability, we introduce Anchor Adherence Rate (AAR), which uses Qwen3-VL to judge whether generated videos satisfy the specified spatial anchoring. AAR is the percentage of videos with correct predicted anchoring. See Appendix B.2 for details.

Implementation Details. We use Wan2.2-TI2V-5B (Wan et al., 2025) as the base model for VFXDirector. We introduce low rank matrices (Hu et al., 2022) with a rank of 32 into the DiT layers. To reduce the training cost, we use center crop and resize videos to 832×480 pixels and evenly sample 49 frames and make sure the anchor points are still valid after adjustment. We use the AdamW (Loshchilov & Hutter, 2019) optimizer with a learning rate of 1e-4. The model is trained on 4 NVIDIA 5090 GPUs for around 32k iterations. During inference, we use 50 denoising steps. See Appendix B.1 for extended details.

5.2 Quantitative Results

The results are presented in Table 1. Video generation models such as CogVideoX and Wan2.2 show limited performance in generating high-quality VFX videos with accurate camera motion and spatial anchoring. Wan2.2-FT shows improved performance over the original Wan2.2, which demonstrates the effectiveness of fine-tuning on our dataset. VFXDirector further improves upon Wan2.2-FT, consistently outperforms the baselines across most metrics by achieving the lowest FVD scores and camera motion errors and the highest AAR, demonstrating the effectiveness of our framework in leveraging both camera and spatial anchoring conditions for unified controllable VFX generation. The results show the ability of VFXDirector to produce visually appealing VFX videos that follow the desired effects and placements under the specified camera motion and spatial anchoring conditions.

Table 1: Quantitative comparison. Best results are in **bold** and second-best results are underlined.

Method	Overall Quality				Camera			Anchor
	FVD↓	Dynamic↑	Subject↑	Imaging↑	CamMC↓	RotErr↓	TransErr↓	AAR↑
CogVideoX	654.07	55.02	69.42	44.69	35.91	5.59	33.72	20.86
Wan2.2	549.17	41.82	79.60	60.38	30.99	3.05	29.94	56.75
LTX-2	417.60	43.59	80.65	56.88	30.33	4.34	28.38	62.82
LingBot-Video	650.48	40.29	77.44	56.38	33.05	3.46	31.87	56.46
Wan2.2-FT	222.08	**64.88**	82.02	62.28	22.66	3.59	21.32	64.78
VFXDirector (Ours)	**197.40**	60.00	**82.45**	**62.45**	**20.62**	**3.00**	**19.48**	**65.55**

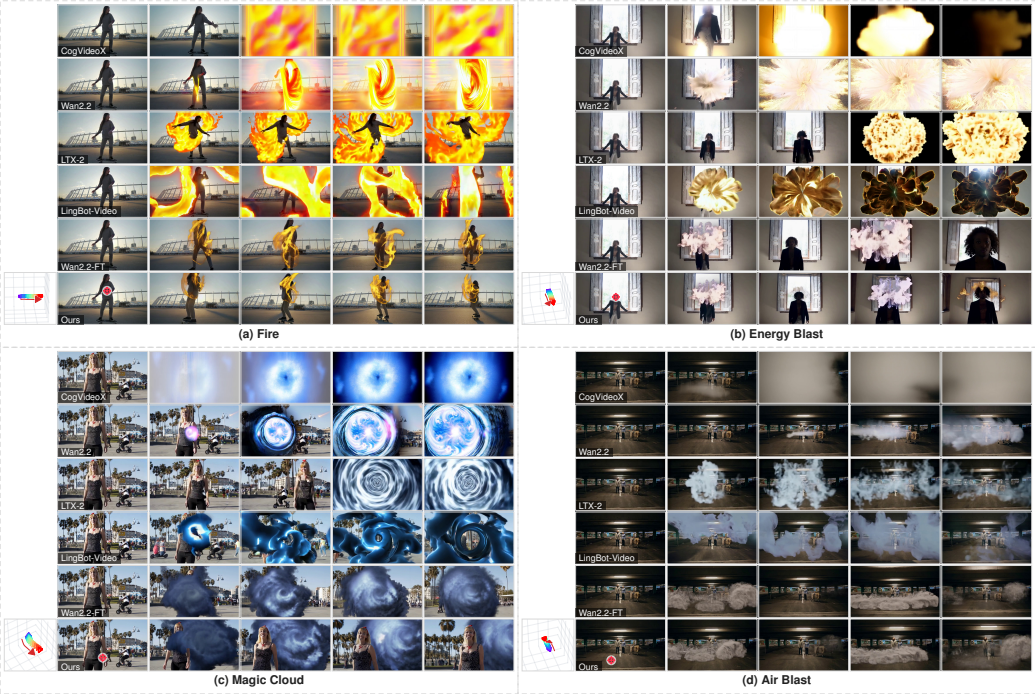

Figure 4: **Qualitative comparison results with baselines on VFXShot.** The first row shows the human-anchored effects (i.e., body and head). The second row shows the scene-anchored effects.

5.3 Qualitative Results

We present qualitative comparisons between VFXDirector and the baselines for unified VFX generation in Fig. 4. Our method can generate high-quality VFX videos with accurate camera motion and spatial anchoring, while the baseline models struggle to generate the desired effects or fail to maintain the correct camera motion and spatial anchoring. The results demonstrate the effectiveness of our framework in generating visually appealing VFX videos that adhere to the specified conditions.

5.4 Ablation Studies

Control Conditions. To validate the contribution of each control condition, we conduct ablation studies by removing either the camera or spatial anchoring conditions from the model. As shown in Table 2, removing camera conditioning consistently increases CamMC, RotErr, and TransErr, demonstrating its contribution to camera-trajectory control. Removing spatial anchoring reduces AAR from 65.55 to 56.70, demonstrating its contribution to maintaining the specified effect placement. The ablations also demenstrate a trade-off between the two control conditions as the full model maintains strong anchor adherence while obtaining the best CamMC and TransErr.

Table 2: Ablation results. Best results are in **bold** and second-best results are underlined.

Variant	Overall Quality				Camera			Anchor
	FVD↓	Dynamic↑	Subject↑	Imaging↑	CamMC↓	RotErr↓	TransErr↓	AAR↑
Input conditions								
Wan2.2	549.17	41.82	79.60	60.38	30.99	3.05	29.94	56.75
w/o Camera	198.65	60.53	**82.58**	**62.58**	23.45	3.84	21.96	**67.46**
w/o Anchor	263.67	**63.44**	81.08	61.51	21.10	2.99	19.95	56.70
Fusion design								
Additive Fusion	204.17	57.42	82.53	62.29	20.97	**2.81**	19.95	64.74
VFXDirector (Ours)	**197.40**	60.00	82.45	62.45	**20.62**	3.00	**19.48**	65.55

Fusion Design. To verify our joint adaptive fusion design, we conduct ablation studies by replacing it with a simple additive fusion. The results are presented in Table 2. The joint adaptive fusion design outperforms the additive fusion on most metrics, which supports its effectiveness in leveraging both camera and spatial anchoring conditions for unified controllable VFX generation.

5.5 User Study

We conduct a user study with 30 randomly chosen participants. Each participant compares videos generated by VFXDirector and the baselines under the same input image, effect description, camera trajectory, and anchoring condition. The videos are presented in randomized order, and participants select the best result according to four criteria: (i) visual quality, (ii) effect fidelity, (iii) camera motion alignment, and (iv) spatial anchoring alignment. As shown in Fig. 5, participants consistently prefer VFXDirector across all four criteria, demonstrating its ability to generate high-quality VFX videos with accurate camera motion and spatial anchoring.

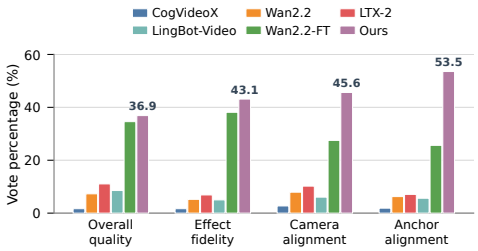

Figure 5: **Results of the user study.** Our VFXDirector is preferred over the baselines.

6 Conclusion

We present VFXShot and VFXDirector for effect-faithful geometry-grounded VFX generation. We propose a dataset construction pipeline to composite VFX assets onto real-world videos with recovered geometry and semantic information. Based on this pipeline, we construct VFXShot, which has 10,298 VFX videos that preserve natural camera, scene, and human dynamics. Built on this data, we introduce VFXDirector, a single video generation model to control the desired effect, camera trajectory, and spatial anchoring. To coordinate camera motion and effect placement, VFXDirector jointly encodes camera and anchor conditions and integrates them into the video DiT through adaptive feature modulation. Our experiments demonstrate the value of the constructed data for VFX generation and show effective controllability.

Limitations. Despite the promising results, our work has several limitations. First, while VFXShot consists of 19 pre-selected VFX assets and two spatial anchoring settings, there is still a need for extending the dataset to broader effect categories, arbitrary objects, and more complex spatial interactions to further increase its coverage. In addition, the geometric information used for data construction is estimated from monocular videos rather than captured by physical sensors. Future work can combine source data with better geometric estimation to support more diverse VFX creation scenarios.

AI USE STATEMENT

We used Qwen3-VL to generate textual descriptions of source videos and VFX assets for dataset annotation and to produce binary adherence judgments used to compute the Anchor Adherence Rate. Apart from these uses, generative AI was used only to improve the readability of the manuscript. We did not use generative AI to construct the composite training videos, formulate theoretical claims, design the methodology or experiments, implement the proposed method, interpret the experimental results, or draw conclusions. All AI-assisted outputs were reviewed by the authors. The authors verified the video descriptions, AAR evaluation protocol and outputs, experimental numbers, citations, figure descriptions, and final manuscript text, and take responsibility for the complete content of this submission.

ETHICS STATEMENT

This work uses source videos from Pexels and animated VFX assets collected from online resources to construct VFXShot through deterministic compositing. The original materials remain subject to their respective licenses and terms of use. We conduct a user study in which participants evaluate anonymized model outputs, and report only aggregate preference rates. The responses are used solely for evaluation and are not included in model training. The proposed method is not intended for surveillance or other abusive or unlawful applications. Applications of the proposed method should respect copyright, privacy, and consent and disclose generated or edited content when appropriate.

REPRODUCIBILITY STATEMENT

The main paper describes the VFXShot construction pipeline, the VFXDirector architecture, and the experimental protocol in Sections 3 to 5. The appendix covers data collection (Appendix A.1), VFX synthesis (Appendix A.2), dataset statistics and preparation (Appendices A.3 and A.4), training and inference (Appendix B.1), metrics (Appendix B.2), and the user study (Appendix B.3). The accompanying implementation is available in the anonymous repository. Together, these materials specify the data pipeline, model training, inference, and evaluation used to obtain the reported results.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and others. Recammaster: Camera-Controlled Generative Rendering From a Single Video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and others. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025b.

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.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, and others. SAM 3: Segment Anything with Concepts. In *International Conference on Learning Representations*, 2026.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, SiYu Zhou, and Qian He. I2VControl-Camera: Precise Video Camera Control with Adjustable Motion Strength. In *International Conference on Learning Representations*, 2025.

Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3DTrajMaster: Mastering 3D Trajectory for Multi-Entity Motion in Video Generation. In *International Conference on Learning Representations*, 2025.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, and others. Motion Prompting: Controlling Video Generation with Motion Trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jiaqi Guo, Lianli Gao, Junchen Zhu, Jiaxin Zhang, Siyang Li, and Jingkuan Song. MagicVFX: Visual Effects Synthesis in Just Minutes. In *Proceedings of the ACM International Conference on Multimedia*, 2024a.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *International Conference on Learning Representations*, 2024b.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, and others. LTX-2: Efficient Joint Audio-Visual Foundation Model. *arXiv preprint arXiv:2601.03233*, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling Camera Control for Video Diffusion Models. In *International Conference on Learning Representations*, 2025a.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. CameraCtrl II: Dynamic Scene Exploration via Camera-Controlled Video Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video Diffusion Models. In *Advances in Neural Information Processing Systems*, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and others. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-One Video Creation and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal Flow Matching for Efficient Video Generative Modeling. In *International Conference on Learning Representations*, 2025.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models. *arXiv preprint arXiv:2606.03748*, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, and others. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*, 2024.

BaoLu Li, Yiming Zhang, Qinghe Wang, Liqian Ma, Xiaoyu Shi, Xintao Wang, Pengfei Wan, Zhenfei Yin, Yunzhi Zhuge, Huchuan Lu, and others. VFXMaster: Unlocking Dynamic Visual Effect Generation via In-Context Learning. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 2026.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, and others. Depth Anything 3: Recovering the Visual Space from Any Views. In *International Conference on Learning Representations*, 2026.

Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, and others. STIV: Scalable Text and Image Conditioned Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

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.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Shuailei Ma, Jiaqi Liao, Xinyang Wang, Jingjing Wang, Chaoran Feng, Zijing Hu, Chong Bao, Zichen Xi, Yuqi Gan, Weisen Wang, and others. Scaling Mixture-of-Experts Video Pretraining for Embodied Intelligence. *arXiv preprint arXiv:2607.07675*, 2026.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent Diffusion Transformer for Video Generation. *Transactions on Machine Learning Research*, 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.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, and others. Motion-I2V: Consistent and Controllable Image-to-Video Generation with Explicit Motion Modeling. In *ACM SIGGRAPH Conference Papers*, 2024.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. MotionStream: Real-Time Video Generation with Interactive Motion Controls. In *International Conference on Learning Representations*, 2026.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, and others. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *International Conference on Learning Representations*, 2023.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards Accurate Generative Models of Video: A New Metric & Challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, and others. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025.

Tun Wang, Hao Sheng, Rongshan Chen, Da Yang, Zhenglong Cui, Sizhe Wang, Ruixuan Cong, and Mingyuan Zhao. Light field depth estimation: A comprehensive survey from principles to future. *High-Confidence Computing*, 4(1):100187, 2024a.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional Video Synthesis with Motion Controllability. In *Advances in Neural Information Processing Systems*, 2023.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A Unified and Flexible Motion Controller for Video Generation. In *ACM SIGGRAPH Conference Papers*, 2024b.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. DragAnything: Motion Control for Anything Using Entity Representation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *European Conference on Computer Vision*, 2024.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. DynamiCrafter: Animating Open-Domain Images with Video Diffusion Priors. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *European Conference on Computer Vision*, 2024.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-Video: Customized Video Generation with User-Directed Camera Movement and Object Motion. In *ACM SIGGRAPH Conference Papers*, 2024.

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*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, and others. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *International Conference on Learning Representations*, 2025.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. DragNUWA: Fine-grained Control in Video Generation by Integrating Text, Image, and Trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E. Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. ReCapture: Generative Video Camera Controls for User-Provided Videos using Masked Video Fine-Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Zhenghao Zhang, Junchao Liao, Menghao Li, ZuoZhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented Diffusion Transformer for Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Mingyuan Zhao, Hao Sheng, Da Yang, Sizhe Wang, Ruixuan Cong, Zhenglong Cui, Rongshan Chen, Tun Wang, Shuai Wang, Yang Huang, and others. A survey for light field super-resolution. *High-Confidence Computing*, 4(1):100206, 2024.

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*, 2026.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. CamI2V: Camera-Controlled Image-to-Video Diffusion Model. *arXiv preprint arXiv:2410.15957*, 2024.

A Dataset Details

A.1 Data Collection

Source Data. To enhance the diversity of our dataset, we collect real-world source videos from Pexels. We search scene-related keywords (e.g. street, road, forest path, hallway) and human-related keywords (e.g. walking, running, raising hand) for videos of various indoor and outdoor scenes with camera motion. Candidate videos are manually checked according to visual quality and their suitability for reliable VFX composition. Specifically, we require: (i) Clear separation between the main subject and background. The subject should not be lost in a cluttered background. (ii) Suitable locations for special effects. At least one location is suitable for adding special effects (e.g., hands, head, around the body, feet, the ground in front of the character, or distant space). (iii) Suitable visual style for presentation. Cinematic, clean, wide shot, medium shot, and dramatic lighting are preferred.

For scene-anchored data, we retain videos with clean scene structure, stable camera motion, and spatial regions that support consistent VFX placement across the shot. For human-anchored data, we use a human pose estimator (Jocher et al., 2026) to select videos containing clearly visible and trackable people, and filter clips with insufficient visibility of the target body keypoints.

Effect Assets. We collect a set of 19 animated VFX assets from online resources which covers visually distinct effect types including fire, smoke, water, lightning, and magic. Each asset is provided as a short video with an alpha channel for compositing. The assets are selected to provide diverse visual appearances and motion patterns, while also being compatible with the source videos in terms of scale, perspective, and scene content. Assets shorter than a source video are looped during synthesis.

Text Annotation. We use Qwen3-VL-8B-Instruct (Bai et al., 2025b) to generate text descriptions for both the source videos and VFX assets, yielding separate scene and effect semantics that can later be composed into generation prompts. Together, the collected footage and VFX assets provide diverse combinations of real camera motion, scene content, human motion, and effect appearance for large-scale synthesis.

A.2 VFX Synthesis Details

Geometric preprocessing. The source videos are normalized to 1920×1080, clipped to at most six seconds, decoded at 24 fps, and resized to 1280×720 for synthesis. For every frame, DA3-GIANT-1.1 (Lin et al., 2026) estimates dense depth and confidence maps, camera intrinsics, and an OpenCV world-to-camera transform. These quantities support Scene-Anchored effects fixed to a 3D scene location and Human-Anchored effects that follow a semantic body location.

Scene-Anchored construction. Let r be a reference frame and $\widetilde{\mathbf{p}}_r = (u_r, v_r, 1)^\top$ a selected 2D anchor with depth d_r. We obtain its world-space position by

$$\mathbf{X}_c^r = d_r K_r^{-1} \widetilde{\mathbf{p}}_r, \qquad \begin{bmatrix} \mathbf{X}_w \\ 1 \end{bmatrix} = (T_{cw}^r)^{-1} \begin{bmatrix} \mathbf{X}_c^r \\ 1 \end{bmatrix}. \tag{3}$$

SAM3 (Carion et al., 2026) identifies likely support and avoid regions on the first frame. Candidate anchors are ranked using semantic support, local depth continuity, geometric confidence, projected effect size, and cross-frame visibility. The selected world point is reprojected through the recovered camera trajectory, keeping the effect attached to the same scene location as the viewpoint changes.

Human-Anchored construction. We use a YOLO-pose (Jocher et al., 2026) estimator with the COCO 17-keypoint convention to track the primary subject. The body anchor is the center of the shoulders and hips, the head anchor is the midpoint of the eyes, the left- and right-hand anchors are the corresponding wrists, and the feet anchor is the midpoint of the ankles. Let $\widetilde{\mathbf{p}}_t^q$ be the homogeneous 2D position of semantic anchor q in frame t, and let $\bar{d}_t$ be the smoothed torso-depth estimate. Its dynamic world-space trajectory is

$$\begin{bmatrix} \mathbf{X}_{w,t}^q \\ 1 \end{bmatrix} = \left(T_{cw}^t\right)^{-1} \begin{bmatrix} \bar{d}_t K_t^{-1} \widetilde{\mathbf{p}}_t^q \\ 1 \end{bmatrix}. \tag{4}$$

Table 3: VFXShot statistics by anchoring mode.

Mode	Source videos	Synthesized videos	Fraction
Scene-Anchored	301	5,719	55.54%
Human-Anchored	241	4,579	44.46%
Total	542	10,298	100.00%

Table 4: Distribution of Human-Anchored relations. "Human %" uses the 4,579 Human-Anchored videos as denominator, and "All %" uses all 10,298 videos.

Anchor	Train	Test	Total	Human %	All %
Body	887	224	1,111	24.26	10.79
Head	880	164	1,044	22.80	10.14
Left hand	789	229	1,018	22.23	9.89
Right hand	687	174	861	18.80	8.36
Feet	405	140	545	11.90	5.29
Total	3,648	931	4,579	100.00	44.46

Unlike the fixed world point in Eq. (3), this point changes over time to follow the person's motion. Short tracking gaps are interpolated, while incomplete or invalid trajectories are removed.

Effect geometry and compositing. Each RGBA asset is represented by a camera-facing 3D quadrilateral whose aspect ratio matches the asset. Its scale is set relative to the image for Scene-Anchored effects and relative to the person's bounding box for Human-Anchored effects. The quadrilateral is projected in each frame, and a homography warps the asset to the resulting footprint. For Scene-Anchored examples, SAM3 foreground masks and depth identify people in front of the effect, and the corresponding effect pixels are suppressed before alpha compositing. Human-Anchored examples are composited without this scene-foreground occlusion step.

Training prompts. The source and effect descriptions are combined with an anchoring instruction. For a Human-Anchored sample, the prompt is:

```
[source video description] Add this visual effect attached
to the person's [anchor label]:  [effect description].
```

For a Scene-Anchored sample, the prompt is:

```
[source video description] Add this visual effect fixed at
the provided scene anchor:  [effect description].
```

Quality control. Scene candidates must satisfy the semantic, geometric, and cross-frame visibility checks described above, while Human-Anchored trajectories must remain valid and continuous throughout the clip. We manually inspect the resulting composites and anchor behavior before freezing the dataset.

A.3 Dataset Statistics

As shown in Table 3 and Table 4, VFXShot contains 10,298 synthesized videos constructed from 542 source videos across the two anchoring modes. These comprise 509 unique Pexels videos, since 33 source videos are used in both modes.

A.4 Training Data Preparation

To reduce the training cost, we construct a dataset package from VFXShot. Each example contains a clean first-frame image, a synthesized target video, the combined text prompt, a normalized 2D

Table 5: Training and inference configuration of VFXDirector.

Setting	Value
Backbone	Wan2.2-TI2V-5B
Input/target resolution	832×480
Frames and output rate	49 frames, $49/6 \approx 8.1667$ fps
Control bottleneck	64 dimensions
Optimizer	AdamW
First-stage learning rate	10^{-4}
Second-stage learning rates	10^{-4} for control modules and 10^{-5} for LoRA
Schedule	constant, no warmup
Weight decay / gradient clipping	0.01 / 1.0
Precision	bfloat16
Per-GPU / global batch size	1 / 4
Gradient accumulation	1
Second-stage duration	8 epochs, 16,416 optimizer steps
Memory setting	gradient checkpointing enabled
Training seed	20260810
Inference sampler	Wan flow-matching scheduler, shift 5.0
Inference steps / text guidance	50 / 5.0
Camera / anchor control scales	1.0 / 1.0

anchor, and per-frame camera matrices. We crop 16 pixels from each horizontal side of the 1280×720 synthesized video and resize it to 832×480, applying the same spatial transformation to the anchor and camera intrinsics. We then sample 49 frames uniformly over the full sequence, using identical indices for the video and camera metadata. We form an 80/20 train-test split with seed 42 by grouping examples according to their deduplicated source video IDs. All 19 effects and both anchoring-mode uses of the same source remain in one partition, so the train and test sets share no source videos.

B Additional Implementation Details

B.1 Training and Inference Details

Backbone and camera representation. VFXDirector uses Wan2.2-TI2V-5B (Wan et al., 2025) as its latent video diffusion backbone. Each 49-frame camera trajectory is converted to first-frame-relative poses and sampled at the 13 temporal latent positions, producing a 13×12 camera input.

Anchor representation. The anchor branch combines a localized spatial signal at the normalized 2D anchor with local appearance sampled from the clean first-frame VAE latent. A 3D convolution maps this representation to 64-dimensional token-aligned features. The text prompt provides the anchoring relation semantics, while the anchor branch provides the spatial location and local appearance.

Joint Adaptive Fusion. At each transformer block, the camera and anchor features are projected into a shared 64-dimensional bottleneck. Joint Adaptive Fusion combines the two conditions and modulates the timestep-conditioned self-attention input through adaptive normalization. Camera and anchor control scales adjust the strength of the two conditions at inference time.

Two-stage optimization. Training proceeds in two stages. In the first stage, we adapt the frozen Wan backbone to the synthesized VFX domain using a rank-32 low-rank adaptation (LoRA) module with $\alpha = 32$, applied to the attention projections and feed-forward linear layers. The checkpoint from the eighth epoch initializes the second stage. This stage focuses on learning our control modules: the camera encoder, anchor encoder, and Joint Adaptive Fusion module are freshly initialized and trained together with the LoRA module. The control modules use a learning rate of 10^{-4}, while the LoRA module uses 10^{-5}. The base transformer weights, VAE, and text encoder remain frozen.

```
Evaluate whether the generated VFX adheres to the specified spatial anchor.

# Input
Anchor relation: {ANCHOR_RELATION}; Effect name: {VFX_NAME}; Effect description: {VFX_DESCRIPTION}.
Annotated first frame: The provided first-frame image contains a clearly marked anchor location. The marker indicates the spatial location selected by the user.
Generated video: The provided video is the evaluation target.

# Task
Watch the complete ordered video sequence and determine whether the generated visual effect adheres to BOTH: 1. the spatial location indicated by the anchor in the first frame; and 2. the specified attachment relation over time. The visual effect may emerge gradually, so do not require a fully visible effect in the first frame. Judge the anchor as successful only when the visible VFX is spatially consistent with the marked location and preserves the required relation during the video. Use the following relation-specific rules.

## If Anchor relation = scene_fixed
Interpret the marked point as a selected location in the scene. Judge TRUE when: the VFX is clearly associated with the selected scene location; as the viewpoint or camera changes, the VFX remains visually attached to the same scene location; and the VFX does not behave like a screen-space overlay or drift arbitrarily away from the selected scene content. Do NOT require exact pixel-level reprojection or known metric 3D depth. Judge visual scene-attachment consistency.

## If Anchor relation = [keypoint]
Interpret the marked point as selecting a person and the required attachment relation as that person's [keypoint]. Judge TRUE when: the VFX is attached to the [keypoint] of the person indicated by the anchor; and the VFX continues to move consistently with that person's [keypoint] over time.

# General judging rules
Judge the spatial origin or attachment of the VFX, not the visual centroid of the whole effect; large effects may extend far away from the anchor and can still be correct if their attachment/origin is consistent; temporary partial occlusion does not automatically mean failure if the relation is otherwise clear; do not require the generated person's motion to exactly match another reference video; do not judge the camera trajectory itself; do not judge the visual quality of the VFX except when severe artifacts make the attachment relation impossible to identify; if the requested VFX never visibly appears, anchor adherence must be FALSE; if the VFX appears but is attached to a clearly wrong scene location, wrong person, or wrong body region, anchor adherence must be FALSE; if the VFX noticeably drifts away from the required relation for a substantial portion of the video, anchor adherence must be FALSE.

# Output
Return exactly this JSON structure:

{
  "anchor_relation": "{ANCHOR_RELATION}",
  "anchor_adheres": true / false,
  "reason": "One concise sentence explaining the spatial and temporal adherence decision in no more than 25 words."
}
```

Figure 6: Prompt template for Anchor Adherence Rate (AAR) evaluation.

We use the original Wan TI2V flow-matching weighted mean squared error (MSE) objective. The complete training and inference configuration is summarized in Table 5.

B.2 METRICS

Video Quality. We report Fréchet Video Distance (FVD) (Unterthiner et al., 2018) using the StyleGAN-V and VideoGPT implementations. FVD measures the distributional distance between generated and reference videos in a spatiotemporal feature space, capturing both visual appearance and temporal dynamics. We additionally use VBench (Huang et al., 2024) to evaluate imaging quality, subject consistency, dynamic degree.

Camera Motion. We evaluate camera control using CamMC (Wang et al., 2024b), RotErr, and TransErr (He et al., 2025a). CamMC measures the discrepancy between the requested and generated camera motion. RotErr and TransErr compare the rotation and translation components of the camera trajectory estimated from the generated video against the target trajectory. Lower values indicate closer alignment with the specified camera motion.

Anchor Adherence. Anchor Adherence Rate (AAR) measures whether the generated effect follows both the initial anchor location and the specified anchoring relation throughout the video. For each sample, Qwen3-VL (Bai et al., 2025b) receives the anchor-marked first frame, effect name and description, anchoring relation, and generated video. As illustrated in Fig. 6, the evaluator applies relation-specific criteria: a scene-anchored effect should remain attached to the same scene location as the viewpoint changes, while a human-anchored effect should follow the specified body region over time. The evaluator returns a binary adherence decision, and AAR is the percentage of generated videos judged to satisfy the requested spatial relation.

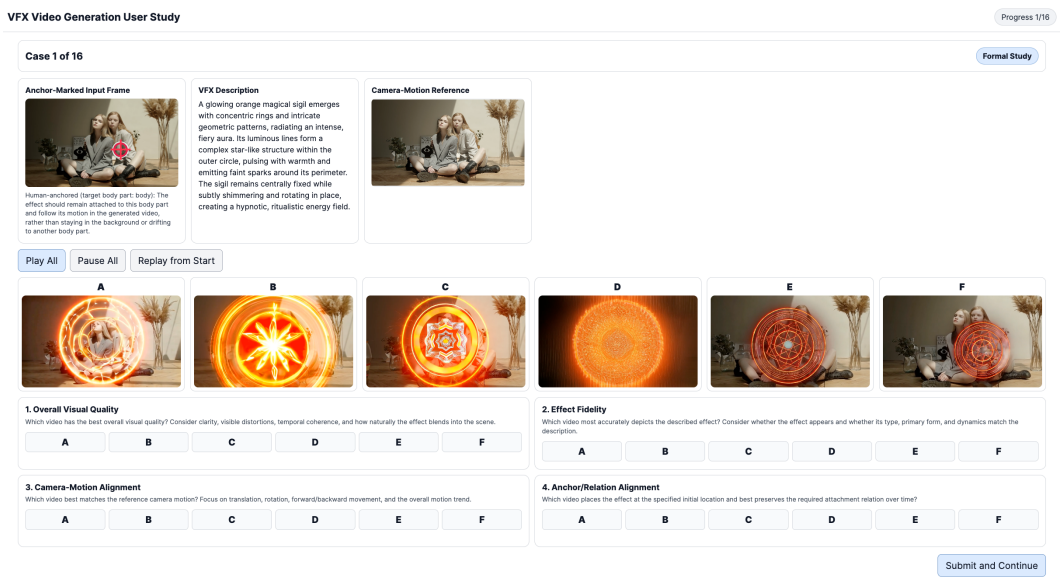

Figure 7: User-study interface. Each case presents the input conditions and anonymized candidate videos, followed by four independent preference questions.

B.3 User Study

We conduct the user study with 30 participants. We randomly select 16 representative cases covering different effects and anchoring types. Each case presents an anchor-marked input frame, the effect description, a camera-motion reference, and anonymized candidate videos generated by the evaluated methods. Candidate identities are represented by letter labels and their display order is randomized. Participants can play the videos together or replay them from the beginning to compare their temporal behavior under the same input conditions.

For each case, participants independently select the best candidate according to four criteria: overall visual quality, effect fidelity, camera-motion alignment, and anchor/relation alignment. Overall visual quality considers clarity, temporal coherence, visible artifacts, and natural integration with the scene. Effect fidelity measures how well the generated effect matches the requested type, appearance, and dynamics. Camera-motion alignment evaluates agreement with the reference camera movement, while anchor/relation alignment evaluates whether the effect appears at the specified initial location and preserves the required scene- or human-anchored relation over time. We compute each method's preference rate as the proportion of selections it receives across all participants and cases. An example of the evaluation interface is shown in Fig. 7.

C Additional Results

We provide additional qualitative comparisons across eight test cases in Fig. 8. Given the same input image, effect description, camera trajectory, and spatial anchor, VFXDirector generates effects that follow the requested camera motion and remain attached to the specified scene location or human body region. The baseline methods more frequently exhibit effect distortion, spatial drift, or inconsistent camera motion. These examples complement the main-paper results and demonstrate consistent control across diverse scenes, effects, and anchoring relations.

D Failure Cases

While our proposed framework has proven effective, certain limitations and failure cases still need to be addressed. First, stationary-camera footage constitutes only a small portion of the real-world source videos used to construct VFXShot. The model therefore receives limited supervision for near-static trajectories and may introduce unintended camera motion even when the target camera

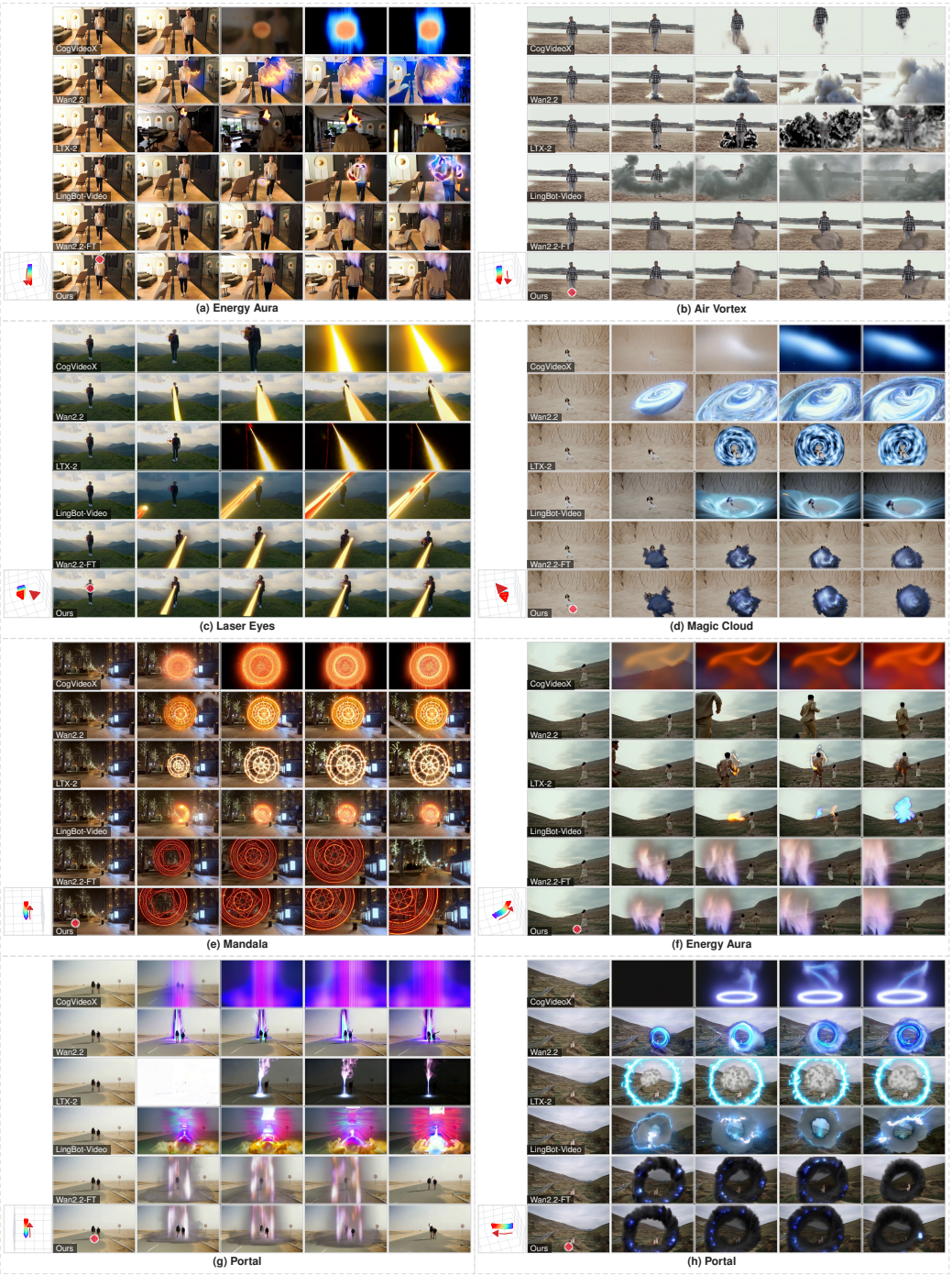

Figure 8: Additional Comparison Screenshot.

remains fixed, as shown in Fig. 9(a). Second, the capacity of the base video-generation model limits its rendering of semi-transparent effects and complex foreground–background interactions. In Fig. 9(b), people or objects that should remain visible through transparent effect regions instead become blurred or obscured, weakening the intended transparent appearance. Future work will expand the coverage of stationary-camera footage and adopt stronger video-generation backbones with improved transparency modeling.

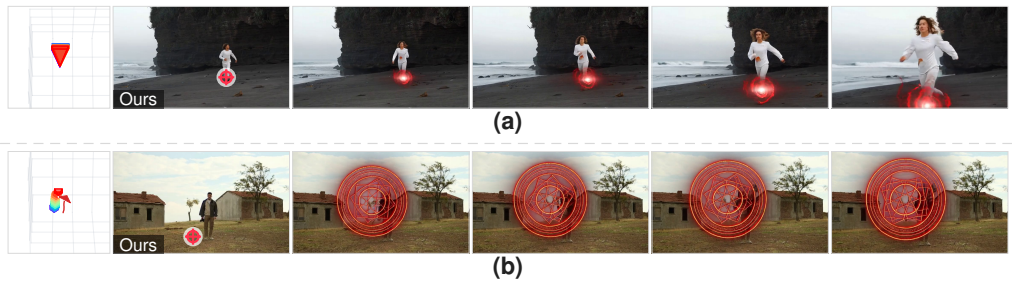

Figure 9: Failure cases. (a) The target trajectory specifies a stationary camera, but the generated video introduces unintended camera motion. (b) The generated effect blurs and partially obscures the underlying content, weakening the intended transparent appearance.