Tuning-free Visual Effect Transfer across Videos

Maxwell Jones[1] Rameen Abdal[2] Or Patashnik[2]
Ruslan Salakhutdinov[1] Sergey Tulyakov[2] Jun-Yan Zhu[1] Kuan-Chieh Jackson Wang[2]
[1]Carnegie Mellon University [2]Snap Research

[1] Carnegie Mellon University
[2] Snap Research

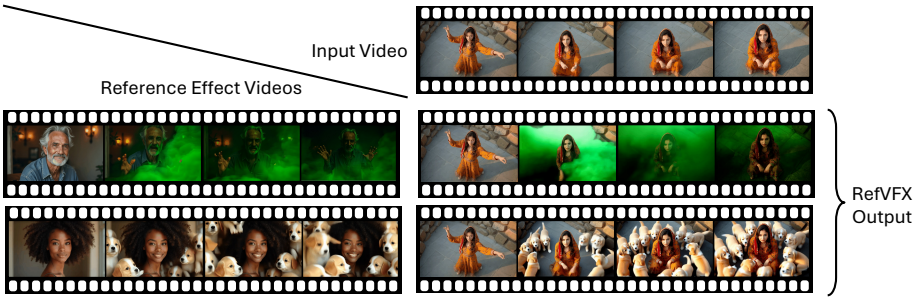

Fig. 1: Overview. We present `RefVFX`, a tuning-free framework for visual effect transfer across videos. Given a reference effect video and an input video, our method produces a new output video where the reference's temporal effect is seamlessly applied to the input's content and motion. Unlike prompt-based or keyframe-conditioned approaches, `RefVFX` directly learns to interpret and reproduce complex time-varying visual effects such as material transformations, object additions, or atmospheric effects from example videos during training, enabling faithful and coherent visual effect transfer at inference time.

Abstract. We present `RefVFX`, a new framework that transfers complex temporal effects from a reference video onto a target video or image in a feed-forward manner. While existing methods excel at prompt-based or keyframe-conditioned editing, they struggle with dynamic temporal effects such as dynamic lighting changes or character transformations, which are difficult to describe via text or static conditions. Transferring a video effect is challenging, as the model must integrate the new temporal dynamics with the input video's existing motion and appearance. To address this, we introduce a large-scale dataset of triplets, where each triplet consists of a reference effect video, an input image or video, and a corresponding output video depicting the transferred effect. Creating this data is non-trivial, especially the video-to-video effect triplets, which do not exist naturally. To generate these, we propose a scalable automated pipeline that creates high-quality paired videos designed to preserve the input's motion and structure while transforming it based on some fixed, repeatable effect. We then augment this data with image-to-video effects derived from LoRA adapters and code-based temporal effects generated

through programmatic composition. Building on our new dataset, we train our reference-conditioned model using recent text-to-video backbones. Experimental results demonstrate that `RefVFX` produces visually consistent and temporally coherent edits, generalizes across unseen effect categories, and outperforms prompt-only baselines in both quantitative metrics and human preference. See our website at this URL

1 Introduction

Generative models have enabled the automatic synthesis and manipulation of images and videos with remarkable realism and diversity [4, 16, 21, 29, 36, 41, 55, 56, 60, 70]. Recent advances in video generation have expanded user control, allowing video editing via text prompts, keyframes, or depth maps [34, 70, 84]. However, most models focus on semantic edits such as modifying objects [56, 85], scenes [56, 85], or styles [46, 79, 85], while overlooking *"temporal effects"* (see Figure 1): effects that *unfold over time* and define the emotional, stylistic, and cinematic character of a video. Effects such as dynamic lighting, intricate camera movements, or character transformations remain difficult to express through text or current visual inputs.

Reference-based conditioning offers a natural way to specify complex temporal behaviors, letting users convey the desired effect through a reference video that captures subtle cues such as motion rhythm, lighting changes, or stylistic transitions. Transferring such temporal effects, especially from one video to another, i.e., a Ref. Video + Input Video → Output Video setup, is especially challenging. To the best of our knowledge, we are the first to show results on this task.

While reference-based editing has shown promise for images [40, 63, 78], extending it to videos poses unique challenges. First, building a dataset is nontrivial: image-to-video (I2V) effects require consistent triplets in which a single static input produces multiple effect-rich outputs, and video-to-video (V2V) effects demand motion-consistent transformations that change only the temporal effect. Without such data, existing systems may fail to learn how to transfer temporal effects independently of the reference video's content or motion. Second, the model must extract temporal dynamics from the reference and seamlessly integrate them with the target's motion and appearance to produce coherent, high quality results. This requires disentangling the effect itself from the specific content and motion of the reference.

We introduce a reference-based video generation dataset, training method, and final framework of finetuned model along with an updated conditioning mechanism dubbed `RefVFX` that transfers complex temporal effects from a reference video onto a target image or video in a feed-forward manner. Built upon recent text-to-video diffusion backbones [21, 36, 70], our model jointly conditions on three inputs: a reference video that provides the temporal effect, an input image or video that defines the scene content and motion, and a text prompt that offers high-level semantic guidance. Through this conditioning, `RefVFX` learns to integrate the reference's temporal dynamics harmoniously with the input's appearance and motion, producing coherent and visually consistent results.

A key contribution of our work is a large-scale dataset designed specifically for this task. Each sample consists of a triplet: a reference video depicting a temporal effect, an input image or video, and a corresponding output video showing the effect applied to the input. Building such a dataset at scale is challenging, as it requires coherent alignment between the reference effect and the target's content and motion. To this end, we construct a unified data curation pipeline that combines three complementary sources: (1) image-to-video effect data derived from existing LoRA-based models, (2) video-to-video transformations generated through an automated pipeline, and (3) synthetic, code-based temporal effects. Together, these sources yield over 1,700 unique effects and more than 120K triplets, providing unprecedented diversity for training reference-conditioned video editors.

We evaluate `RefVFX` across diverse unseen temporal effects, comparing it with recent prompt-based and reference-guided video editing models [25, 34, 70]. Qualitatively, our method produces coherent, visually consistent videos that successfully integrate the transferred effects while preserving the motion and appearance of the input. Quantitatively, `RefVFX` achieves higher scores in reference video embedding similarity to the reference effect video; however, existing metrics only partially reflect the aesthetic and temporal nuances of effect transfer. To better assess perceptual quality, we conduct a human preference study, where participants consistently favor our results, highlighting the importance of human judgment for evaluating dynamic visual effects. Notably, the model generalizes to unseen categories of effects and operates in a feed-forward manner without optimization at inference time, demonstrating both robustness and efficiency. Our code will be released upon publication.
Our main contributions are summarized as follows:

- We introduce `RefVFX`, a framework for reference-based video effect transfer that enables users to apply complex temporal effects from a reference video onto arbitrary videos or images in a feed-forward manner.
- We construct a large-scale, effect-aligned dataset comprising over 120K triplets of (reference, input, output) videos covering more than 1,700 distinct temporal effects along with a validation set, establishing a new benchmark for future research.
- We design a multi-source conditioning architecture built upon recent diffusion backbones that jointly encode reference video dynamics, input appearance/motion, and text prompts.
- We conduct extensive qualitative, quantitative, and human preference evaluations, demonstrating that our method achieves superior visual and temporal coherence compared to prompt-only or reference-guided baselines, and generalizes effectively to unseen effect categories.

2 Related Work

Text-to-Video Generation. Recent advances in text-to-video generation have significantly improved video quality, temporal coherence, and prompt fol-

lowing [4, 13, 21, 26, 28, 29, 36, 50, 56, 70]. Similar to text-to-image generation [16, 41, 55, 60], these models are largely based on diffusion or flow-matching in latent space [3, 15, 45, 49, 66]. Early work adopted U-Net backbones [19, 61, 64], while modern approaches employ diffusion transformers [53] that embed videos into latent patches and apply bidirectional attention with text [21, 26, 29, 36, 56, 70]. Beyond pure text conditioning, some models perform image-to-video or first–last-frame generation [21, 70], where a user provides keyframes and a text prompt to guide synthesis. However, these settings modify only edge frames, whereas our approach conditions on both an entire input video and a reference effect video, enabling transfer of *temporal effects* that evolve throughout the sequence while preserving input motion and appearance.

Reference-Based Controllable Generation. Reference-based conditioning uses a reference image or video that is not spatially aligned with the output. In text-to-image models, it is widely used for identity preservation [11, 20, 39, 43, 51, 52, 63, 68, 73, 75, 78] or style transfer [17, 24, 31, 62, 65, 76, 81]. These signals are injected via per-reference optimization [17, 63, 65], cross-attention [20, 43, 51, 68, 76, 78], or added reference tokens [24, 39, 62]. Video generative models extend these paradigms temporally [10, 22, 32, 44, 46, 47, 71, 79, 80], typically conditioning on one or more *static* reference images (identity or style) alongside text. Such methods maintain consistent appearance but cannot model *temporally evolving* effects. More recent finetuning based methods such as Dynamic Concepts [1, 2] are also inefficient, requiring either a new LoRA per effect instance or trained on limited data hampering scalability. In contrast, `RefVFX` conditions on a *reference video* that encodes dynamic, time-varying effects, allowing generalizable transfer of phenomena such as lighting changes, camera motion, stylistic transitions or identity-based conditioning.

Video Editing. Video editing has progressed through both zero-shot and supervised methods [6–8, 14, 23, 37, 42, 69, 72, 74, 82]. Pretrained text-to-image models can perform zero-shot edits [6, 8, 14, 23, 37, 69], while paired datasets of object or style edits [5, 72] have enabled general-purpose editing models [7, 42, 74]. Video editing methods follow similar patterns: some are zero-shot [9, 18], while others train paired text-driven editors [12, 48, 85]. However, existing methods rely solely on text and an input video, enabling semantic or appearance edits but offering no mechanism to control *how* visual properties evolve over time. Our approach introduces an additional reference video encoding a temporally varying effect, enabling control over dynamic phenomena such as lighting transitions, motion-dependent effects, and stylistic evolution that static, per-frame, or text-only paradigms cannot capture.

3 Method

3.1 Overview

Our goal is to transfer a temporally evolving visual effect from a *reference video* onto a separate *input image or video*. The resulting output should preserve the content and motion of the input while exhibiting the temporal dynamics shown

(a) LoRA I2V Effects (b) Custom V2V Effects (c) Code Based Effects

Fig. 2: Dataset Examples. We show example triplets from each of our datasets. (a) We curate a reference video + image to video dataset by collecting Low Rank Adapters (LoRAs) [30] for different Image to Video effects online. For each effect, we can apply its corresponding LoRA to two separate images to create a triplet. (b) We create a custom pipeline for generating text guided reference video + input video to video effects. For more details, see Fig. 4. (c) We generate a large scale set of (ref video, input video, output video) triplets by curating specific code pipelines that apply specific, detailed effects to arbitrary videos. Armed with a specific code based effect and a fixed set of hyperparameters, we can apply the exact effect to an arbitrary number of input videos to create many triplets.

in the reference. To achieve this, we introduce `RefVFX`, which combines (1) a large-scale effect–aligned triplet training dataset, (2) a pretrained diffusion-based video generation model, and (3) an updated conditioning scheme allowing the video generation model to condition jointly on the reference video, input video, and text.

First, we give a high level overview of our dataset, followed by describing how we create each individual subset of the dataset in detail. Following this, we describe the model architecture for `RefVFX` and implementation details for training.

3.2 Training Data

Our training data consists of triplets of the form
(reference video, input image or video, target video). To create these triplets, we first curate N effects $\{E^i\}_{i=1}^N$ (e.g., rain, motion blur, color shift), with each effect E^i associated with a set S^i of K input output pairs:
$S^i = \{(\text{input}_j^i, \text{output}_j^i)\}_{j=1}^K$. Here, output_j^i is obtained by applying E^i to input_j^i. To form a training triplet, we select an effect E^i and two distinct pairs $(\text{input}_j^i, \text{output}_j^i)$ and $(\text{input}_\ell^i, \text{output}_\ell^i)$. We discard the input of the first pair, and construct triplet $(\text{reference} = \text{output}_j^i,\ \text{input} = \text{input}_\ell^i,\ \text{target} = \text{output}_\ell^i)$, where the model takes in the first two videos and must predict the third. With this objective, the model learns to transfer the temporal effect from the reference onto the input while preserving the input videos motion and content. Triplets are derived from three complementary sources: (1) LoRA-based image to video effects, (2) video to video effects generated through our scalable pipeline, and (3) programmatically defined temporal effects, which we describe below and showcase in Fig. 2.

Image to Video Effects: We curate a set of image-to-video effects using open-source Low-Rank Adapters (LoRAs) [30] trained on small sets of videos sharing

Fig. 3: Dataset Subset. We display a summary of our dataset structure and available sample effects. These include LoRA-based Image-to-Video, our scalable V2V pipeline (see Fig. 4) , and programmatic methods. For code-based effects, individual tasks are created by combining specific effects (e.g., Glow) and transitions (e.g., Fade In) with randomized hyperparameters.

a common visual effect atop a base image-to-video model [70]. Each adapter is treated as a distinct effect E_i. For every effect, we generate a large collection of videos from diverse input images synthesized via high-quality image generation models [42,74], where each per-adapter set forms one effect category. Triplets are composed of a reference video with effect E_i, the first frame of the target, and the corresponding output video. In total, we source 43 LoRAs with an average of 300 videos per effect, yielding over 14K video clips. See Fig. 3 for example LoRAs.

Video to Video Effects: We propose a scalable video-to-video data generation pipeline for a broad class of *motion-based effects*. These effects manifest *over time* rather than being the product of a single frame image edit followed by propagation of the edit to all frames (e.g., "turn the person into metal" as compared to "add a hat to the person throughout"). To our knowledge, this is the first scalable method designed for creating a large-scale training data pipeline for motion-based video editing.

Given an arbitrary effect prompt, our method produces paired videos (V, V'), consisting of an original video V and its edited counterpart V' exhibiting effect E_i (Fig. 4, 3). **Scalable V2V Algorithm.** The process for generating these (V, V') pairs is as follows: We begin by generating a subject image I using a high-quality image generation model [41, 74]. Using an image editing model [42, 74], we modify I to produce an image I' that depicts the same subject under a new pose, camera angle, and facial expression, which will be used to anchor the last frame of our video pair. To obtain the last frame of the edited video with effect E_i, we apply an image editing model on I' [42, 74] to apply effect E_i, resulting in I''. Given the unedited image I and the edited image I', we synthesize video

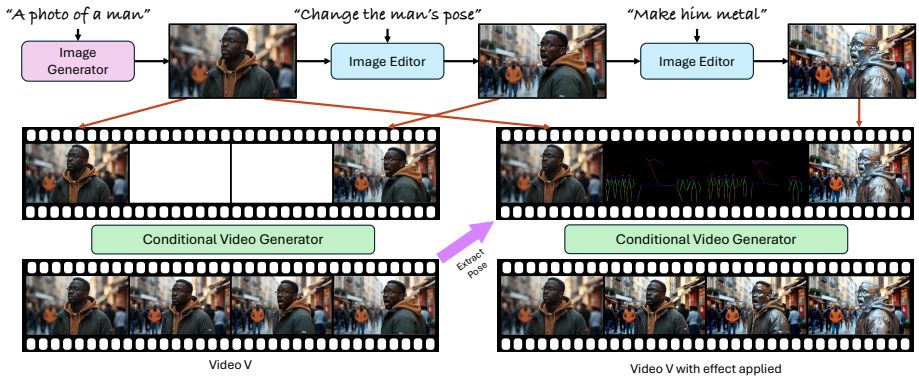

Fig. 4: Text-Guided Video Editing Pair Creation. We present a method to generate a pair of videos (V, V') from an effect prompt E, where V is an initial video and V' is the video with an effect E applied. First, an image generation model is used to create an initial image. Next, an image editing model is used to change the pose, camera angle, and facial expression of the image. Finally, this image is again edited to add effect E. The first and second generated images are used with a first last frame model to output video V. Then, we use a conditional video model conditioned on the original first frame, effect edited last frame, *and intermediate poses from video* V as conditioning to create video V'.

V using a first–last frame interpolation model [70]. Once we have V, we extract intermediate poses using an image-to-pose model [77], which will be used as conditioning for our second video generation. Finally, we generate the edited video V' by leveraging a conditional video generation model [34], which can take in independent conditioning signals for each output video frame generated. We take advantage of this property, and condition the video generation on:

- Initial image I for first frame conditioning
- Final edited image I'' for last frame conditioning
- Intermediate pose images extracted from V for frames 2 through $N-1$

This enables framewise consistency between the input and output video as well as realistic temporal transitions between the unedited initial frame I and stylized final frame I''. For a visual, see Fig. 4.

Effect Generation. To ensure diversity, we automatically generate a large set of motion-based effect prompts using GPT-4o [33]. These effects span multiple semantic and visual categories, which we detail in the supplement.

Programmatic Temporal Effects. We further construct a large-scale dataset of *programmatically defined temporal effects* applied to real videos. Our approach enables scalable and reproducible creation of diverse video effects with fine-grained control over appearance and temporal dynamics.

Base Videos. We source input videos from the Senorita dataset [85], which provides high-quality clips with accompanying foreground–background segmen-

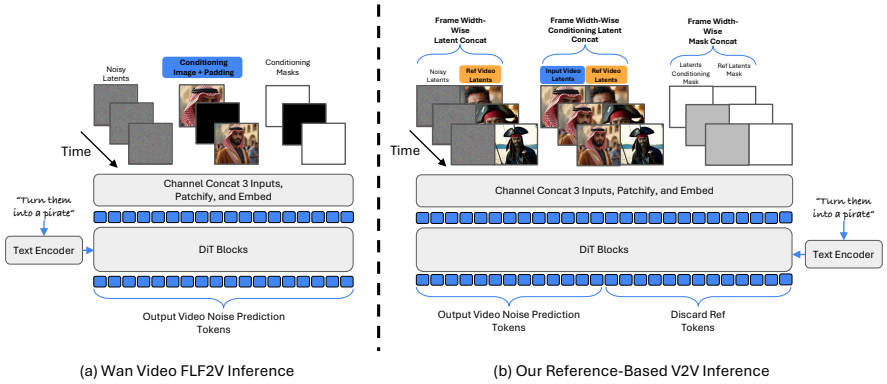

Fig. 5: Architecture overview. (a) Standard Wan Video First–Last Frame to Video (FLF2V) architecture: noisy spatio-temporal latents are channel-wise concatenated with conditioning inputs and a mask, then patchified, embedded, and processed by the diffusion transformer to predict velocity. (b) In our setup, x input video latents are used as conditioning for the noisy latents, while reference video latents are concatenated width-wise to both. The latent mask is set to 1 for frames preserved exactly in the output and 0.5 for those to be modified; the reference latent mask is all ones. This design doubles the token count relative to base generation while conditioning on both the reference effect video and input video. Since all three inputs are channel-concatenated before patchification, repeated clean reference latents are merged channel-wise before embedding, ensuring no redundant reference information across tokens.

tation masks obtained using SegmentAnything [59]. These masks allow effects to be applied selectively to the full frame, the foreground, or the background.

Effect Library. We define a library of code-based effects c_i such as posterization, pixelation, and dithering. Each effect includes hyperparameters that control its visual characteristics. For example, posterization varies in the number of color bins and palette type, while pixelation varies in block size. Sampling different parameter configurations yields a wide range of distinct appearances within each effect class.

Temporal Compositing. To introduce temporal variation, we consider a set of temporal transition operators t_i (e.g., wipe-right, diagonal wipe, circular-out, etc). Each transition is also parameterized by temporal hyperparameters controlling its start time and duration.

Final Composition. For each composite effect E_i, we randomly sample an effect mask (e.g., foreground, background, or all), a spatial effect, its parameters, a temporal transition, and its corresponding temporal parameters. See Fig. 2 (c) for an example. This final effect E_i is then applied to a fixed set of source videos, generating K unique clips per effect. In total, this pipeline yields approximately 100K videos spanning 1,500 distinct synthetic motion-based effects.

3.3 Model Architecture and Conditioning

We build upon the Wan2.1 conditioning model architecture [34, 70], which extends text-to-video diffusion models to image-to-video (I2V), first-last frame to

	Neural V2V			Code-Based V2V		
Method	RVA	IVA	OM	RVA	IVA	OM
Lucy Edit	62.3 ± 2.9	60.7 ± 2.8	65.7 ± 3.1	64.7 ± 4.5	61.4 ± 5.0	65.3 ± 5.0
VACE (Depth)	58.3 ± 2.5	59.1 ± 2.6	66.8 ± 2.1	65.7 ± 5.4	56.2 ± 5.0	61.9 ± 5.0
VACE (Pose)	57.0 ± 2.1	53.7 ± 2.2	60.0 ± 2.2	62.9 ± 4.8	63.3 ± 4.9	64.3 ± 4.7
Ours (V2V Training)	57.8 ± 2.4	53.9 ± 3.1	60.7 ± 3.2	66.2 ± 4.5	61.1 ± 5.0	66.2 ± 4.1
Ours (V2V + I2V Training)	55.2 ± 3.2	52.6 ± 3.5	56.1 ± 2.8	55.5 ± 3.9	56.2 ± 4.0	57.5 ± 3.5
Ours (No Ref)	57.5 ± 2.8	51.3 ± 3.1	67.2 ± 2.9	59.5 ± 3.5	54.3 ± 4.0	61.3 ± 3.9

Table 1: User Study Results for Neural V2V and Code-Based Edits. Each cell shows Win Rate (%) with standard deviation. Our method shows large preference over base models as well as ablations on training dataset and input video(s) in Reference Video Adherence (RVA), Input Video Adherence (IVA), and Overall Match (OM), where we ask users to judge which method best adheres to the input video *while applying the reference video effect.*

video (FLF2V), depth to video, and other conditioning types by concatenating noisy latent channels with conditional and mask channels. We extend this mechanism to jointly condition on both a reference video effect and an input video.

The input video's latents are supplied as conditioning latents, while the reference effect video is encoded into additional latents that are concatenated across all frames. This allows for spatial self-attention between noisy latent tokens with input video conditioning information and reference effect tokens during a forward pass. A hybrid mask controls which latent conditioning frames are preserved or edited based on which are directly copied to the final output and are changeable. The architecture is shown in Fig. 5. Further architectural details, including input condition dropout and cross attention configurations are provided in the supplementary material.

3.4 Implementation Details

We fine-tune the Wan2.1 [70] 14 Billion parameter First-Last Frame to Video model with Low Rank Adapters for 10K steps with a batch size of 8 on a single node of 8 NVIDIA A100 GPUs. We sample from each of the three datasets equally during training, and drop the reference effect video conditioning, control video conditioning, and text prompts with low probability to allow for various forms of classifier-free guidance [7, 27] during inference time. We also add the true last frame from the target video to the model with low probability to allow for the first and last frame capabilities of the underlying model to be retained under our new setup.

4 Experiments

We evaluate our approach on both reference-based image-to-video and reference-based video-to-video generation tasks. In each setting, we provide a pair (reference video, input image or video), and assess the quality of our generated

Baseline	Reference Video Adherence	Overall Match
Wan2.1	57.1 ± 2.4	63.4 ± 2.4
VACE (I2V)	57.4 ± 3.2	62.9 ± 2.6
No Ref	56.8 ± 2.3	61.8 ± 3.4

Table 2: User Study Results for I2V. Each cell shows Win Rate (%) with standard deviation. Our method shows large user preference over baselines in Reference Video Adherence and Overall Match (OM), where we ask users to judge which method best preserves the input image *while applying the reference video effect.*

outputs against strong baselines. Our evaluation emphasizes generalization to unseen reference effects and motion patterns.

4.1 Validation Datasets

Image-to-Video Testing. We construct an evaluation set of unseen image-to-video effects sourced from publicly available LoRA-based visual effect models that were not included in training. For each LoRA effect, we generate multiple reference videos from distinct input images, resulting in 28 unique unseen LoRA effects and a total of 56 reference videos for evaluation.

Video-to-Video Testing. We generate a validation set of Reference Video Effect + Input Video pairs for our reference effect based video to video task. For the first source of data, we reuse the unseen LoRA effects from the previous section as reference effect videos and replace the input images with unseen real input videos to produce the input video pair. Secondly, we use our pipeline introduced in Section 3.2 with novel prompts not seen during training for a portion of the validation dataset. Each resulting pair (reference video, input video) generated using our method is manually filtered to ensure high perceptual quality. Finally, we synthesize an additional set of temporally varying effects using the programmatic procedure from Section 3.2, but with hyperparameter choices unseen during training. We collect over 100 validation instances, and separate the Code-Based Video-to-Video Effects from the Neural Video-to-Video effects when reporting results.

4.2 Baselines

Since existing baselines cannot directly condition on reference videos, we compare our method to state-of-the-art text-conditioned image-to-video and video-to-video models. All baselines receive the same textual descriptions of the desired effects as those implicit in our reference videos.

Image-to-Video Baselines. We benchmark against the Wan 2.1 image-to-video model [70] and the Wan VACE model [34], which represent current state-of-the-art text-guided video generation systems.

Video-to-Video Baselines. For the video-to-video task, we evaluate two configurations based upon the Wan VACE [34] conditional model and one native text-based video editing model. Pose-conditioned Wan VACE: In this setup, we condition the Wan VACE on the first frame of the input video, intermediate

poses extracted from the input video [77], and the corresponding text prompt. Depth-conditioned: In this setup, we condition the Wan VACE on the first frame of the input video, intermediate depth maps [58] from the input video, and the same text prompt. Lucy Edit [25]: Lucy Edit is a text-based video editing model built on Wan 2.2 [25], which directly takes an input video and a text instruction to produce an edited output.

4.3 Ablations

We ablate our method with respect to dataset and input videos. For dataset ablations, we train models for the same number of steps, but with only a subset of the data. We test training only on the Video to Video Effects from Section 3.2, as well as both the Video to Video Effects and LoRA based Image to Video Effects from Section 3.2. To test if our final model actually uses the reference effect video, we also run inference on the final model without the reference effect video, only conditioning on the input video. Since we drop the reference effect video with some percentage during training, the model outputs plausible videos in this setting.

4.4 Qualitative Results

Image-to-Video Effects. Figure 6 shows qualitative comparisons for the image-to-video task using unseen reference effects. Baseline models, which rely solely on text prompts, struggle to capture fine-grained temporal and stylistic cues. In contrast, our method accurately transfers the reference video's camera motion, lighting, and character transformations, producing coherent temporal dynamics that align closely with the target effect.

Video-to-Video Effects. Qualitative comparisons for the video-to-video task are shown in Figure 7. Wan VACE variants either overfit to the input video or introduce artifacts, while Lucy Edit produces static, frame-invariant edits that lack temporal evolution. By conditioning on the full reference video, our method consistently reproduces complex time-varying transformations such as gradual color shifts or structural morphing while maintaining spatial consistency and motion alignment with the input.

User Study For a subjective task such as reference-based video editing or image-to-video generation, it is difficult to determine any single quantitative metric that fully captures success. We therefore conduct a comprehensive user study as the primary evaluation of our method across three tasks: neural video-to-video effect transfer (V2V), image-to-video effect transfer (I2V), and code-based temporal effects. Annotators on Amazon Mechanical Turk performed two-alternative forced choice (2AFC) pairwise comparisons between our method and all baselines and ablations. Each comparison was evaluated along three dimensions: Reference Video Adherence (RVA), Input Video Adherence (IVA), and Overall Match (OM), where OM asks which output best applies the reference effect while preserving the input video. For I2V, we evaluate on RVA and OM. As shown in Table 1, our method achieves consistent preference over all baselines in V2V across all three dimensions, with win rates of 57.0–62.3% in RVA, 53.7–60.7%

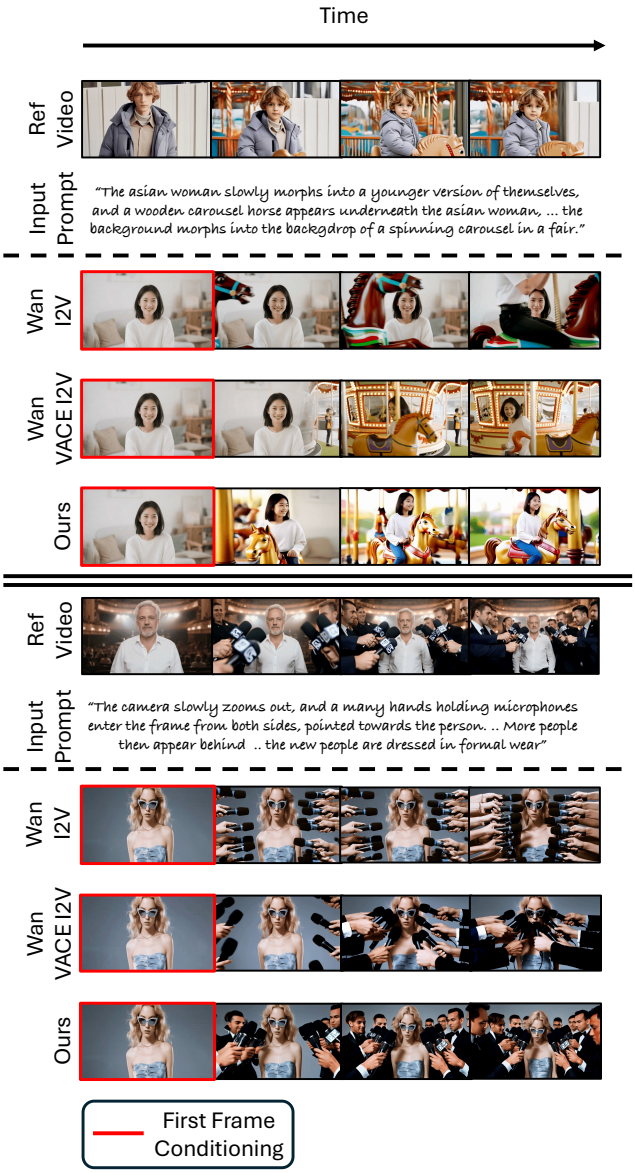

Fig. 6: Qualitative Reference Based Image to Video Results. In the top example, baselines are unable to turn the input woman into a younger version of herself, or put her on the carousel. Our method correctly makes the subject younger, puts her on the carousel, and mimics the camera motion of the reference video. In the bottom example, baselines are unable to add the reporters into the scene, while our method correctly adds hands with microphones followed by reporters, and mimics the interactions and occlusions from the reference video.

in IVA, and 60.0–67.2% in OM. Code-based effects show even stronger results, with win rates of 59.5–65.7% in RVA and 61.3–65.3% in OM. For I2V (Table 2),

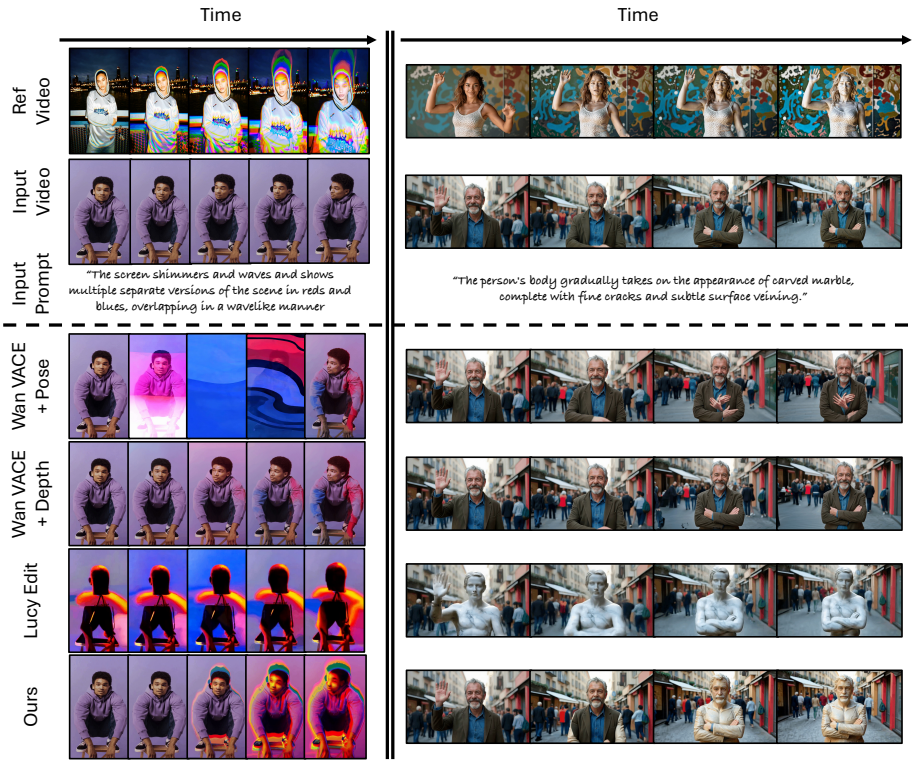

Fig. 7: Results for our reference-based video-to-video setup. We compare three baselines: (1) Wan VACE + Pose: a conditional diffusion transformer [34] using the first input frame and extracted poses; (2) Wan VACE + Depth: using the first frame and extracted depths with Wan VACE; and (3) Lucy Edit [25], a video editing model fine-tuned from an image-to-video model. Without reference video information, all baselines fail to follow the reference from a prompt alone. Wan VACE methods overfit to the input or add irrelevant edits, while Lucy Edit applies uniform static edits across frames. Our method successfully follows the reference effect.

we observe win rates of 56.8–57.4% in RVA and 61.8–63.4% in OM against all baselines. For dataset ablations, we compare against models trained for the same number of steps on subsets of our data: V2V effects only, and V2V + I2V effects (excluding code-based effects). Our full model is preferred over both dataset ablations across all metrics, confirming that training on the complete mixture of effect types is beneficial. Our model trained on all datasets clearly beats ablations on training data for Code-Based V2V effects, which intuitively is clear as only the final model trains on our synthetic code based edits. Notably, our full model also beats dataset baselines on the Neural V2V dataset, indicating that our extra data sources also yield better performance on general video effects from unseen LoRAs. All reported win rates for overall quality exceed the 50% chance level, standard deviations confirm statistical reliability. The closest statistic in

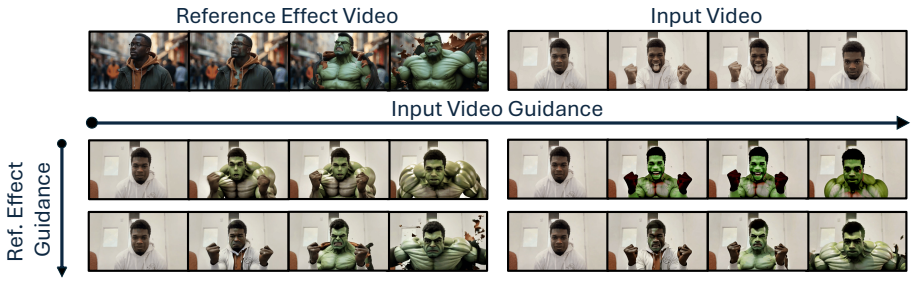

Fig. 8: Qualitative Results for Controllable Real Video Editing. We show an example of the effect of reference effect guidance and input video guidance on model output. The first row shows the reference effect video and input video that are used for model conditioning. From left to right, we increase the input video guidance, while from top to bottom we increase the reference effect guidance. Notice that as input video guidance increases, the input video details are more faithfully kept (i.e. attire, facial shape). As the reference effect guidance increases, the reference effect video takes a larger role (i.e. reference effect muscle shape, and clothing of the man in the reference effect). Applying both gives a mix of both input video and ref. effect video adherence.

terms of preference is the input video adherence for our method without reference input video, which is expected as it uses the same model as our full method, and the test is simply adherence to the input video which both versions take as input. These results validate our approach's effectiveness in effect transfer while preserving input content fidelity across multiple generation modalities. **In total, we collected over 19000 annotations from over 500 unique annotators.** For more details on the user study setup, please refer to the supplement.

5 Controllable Real Video Editing

When editing real videos, users may want to control the level to which the original video is maintained, as well as the amount that the reference effect video effects the final generation. To this end, we follow previous work [7,27,35,38,39,67] and combine multiple classifier-free guidance directions during inference. Specifically, we consider applying guidance for text, input video, and reference effect video, normalizing latents as recommended in previous work [39]. Interpolating between different guidance vectors results in output videos with different amounts of adherence to the reference effect video and input video respectively, as seen in Figure 8. For the full formula, please see the supplement.

5.1 Quantitative Metrics

Conditioning Input Similarity. Since our validation sets do not include ground-truth triplets, we measure the similarity of our generated outputs to both the input and reference videos as a proxy for task success. To quantify this, we employ VideoPrism [83], a large-scale video embedding model pretrained on over 500M clips, to compute feature-space similarities between videos. For the reference + image-to-video case, we instead measure the average similarity

Method	First Frame Sim.	Ref Sim.
Wan 2.1	**0.7911**	0.7230
Wan VACE I2V	0.7799	0.7127
Ours	0.7698	**0.7378**

Table 3: I2V Similarities. First Frame Sim. is average CLIP [57] similarity to the input frame. Ref Sim. is video embedding similarity [83] between the output and reference effect video. **Bold**: best, underline: second best.

Method	Neural V2V		Code-Based V2V	
	Input Sim.	Ref Sim.	Input Sim.	Ref Sim.
Wan VACE Pose	0.9068	0.6539	0.9225	0.6002
Wan VACE Depth	**0.9460**	0.6226	0.9394	0.5998
Lucy Edit	0.7544	0.6852	0.8882	0.6539
Ours	0.8568	**0.7014**	**0.9479**	**0.7169**

Table 4: Neural and Code-Based V2V Similarities. Input Sim. is video embedding similarity [83] between output and input video. Ref Sim. is the same between output and reference effect video. **Bold**: best, underline: second best.

between the generated video and its first-frame conditioning using CLIP [57] embeddings. Results are summarized in Tables 3 and 4. Our method consistently achieves higher similarity to the reference video than all baselines, confirming that it effectively incorporates temporal and stylistic information from the reference. Interestingly, baseline models often exhibit slightly higher similarity to the input, likely reflecting a form of under-editing. This phenomenon is also visible in Figure 7 (left) and Figure 6 (bottom), where baselines preserve input structure but fail to reproduce the desired temporal evolution.

6 Discussion

Misuse and Deepfake Potential. The same expressive capabilities that make `RefVFX` compelling for creative use -also carry well-known risks when applied to identifiable subjects, most notably the production of non-consensual deepfakes or otherwise misleading content. We share these concerns and have taken concrete steps to mitigate them. All real videos used for training are drawn from openly licensed sources, specifically Señorita-2M (CC-NC) [85] and Pexels (free use) [54], with all other images generated from image or video generative models [70, 74]. Alongside our model and dataset release, we will provide a model card that explicitly documents intended use, limitations, the heightened risk of identity-related manipulation and an acceptable-use clause.

Conclusion. In conclusion, we presented `RefVFX`, a tuning-free framework for transferring complex temporal visual effects across videos using reference conditioning. By introducing a large-scale dataset of over 120K triplets encompassing 1,700 distinct effects, we enable the first systematic study of reference-based temporal effect transfer. Our scalable video-to-video generation pipeline produces diverse, motion-consistent training pairs, while our diffusion-based architecture jointly encodes reference dynamics, input appearance, and semantic guidance to produce coherent results. Perceptual and quantitative evaluations demonstrate that `RefVFX` surpasses prompt-only and static reference baselines in both visual fidelity and temporal coherence.

Acknowledgments We would like to thank Yusuf Dalva, Rishubh Parihar, James Burgess, Ruihang Zhang, Sheng-Yu Wang, Gaurav Parmar, and Nupur Kumari for their insightful feedback and input that contributed to the finished work. Maxwell Jones is supported by the Rales Fellowship. This project is partly supported by the Packard Fellowship.

References

1. Abdal, R., Patashnik, O., Deyneka, E., Chen, H., Siarohin, A., Tulyakov, S., Cohen-Or, D., Aberman, K.: Zero-shot dynamic concept personalization with grid-based lora (2025), https://arxiv.org/abs/2507.17963
2. Abdal, R., Patashnik, O., Skorokhodov, I., Menapace, W., Siarohin, A., Tulyakov, S., Cohen-Or, D., Aberman, K.: Dynamic concepts personalization from single videos. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. SIGGRAPH Conference Papers '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3721238.3730644, https://doi.org/10.1145/3721238.3730644
3. Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. arXiv preprint arXiv:2303.08797 (2023)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
5. Boesel, F., Rombach, R.: Improving image editing models with generative data refinement. In: The Second Tiny Papers Track at ICLR 2024 (2024)
6. Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., Passos, A.: Ledits++: Limitless image editing using text-to-image models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8861–8870 (2024)
7. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
8. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
9. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
10. Chen, T.S., Siarohin, A., Menapace, W., Fang, Y., Skorokhodov, I., Zhu, J.Y., Aberman, K., Yang, M.H., Tulyakov, S.: Videoalchemy: Open-set personalization in video generation (2024)
11. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6593–6602 (2024)
12. Cheng, J., Xiao, T., He, T.: Consistent video-to-video transfer using synthetic dataset. arXiv preprint arXiv:2311.00213 (2023)
13. cloneofsimo: fofr/cog-aura-flow. https://github.com/fofr/cog-aura-flow (2024), gitHub repository
14. Deutch, G., Gal, R., Garibi, D., Patashnik, O., Cohen-Or, D.: Turboedit: Text-based image editing using few-step diffusion models. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–12 (2024)
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)

16. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
17. Gao, J., Sun, Y., Liu, Y., Tang, Y., Zeng, Y., Qi, D., Chen, K., Zhao, C.: Styleshot: A snapshot on any style. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
18. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
19. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
20. Guo, Z., Wu, Y., Zhuowei, C., Zhang, P., He, Q., et al.: Pulid: Pure and lightning id customization via contrastive alignment. Advances in neural information processing systems **37**, 36777–36804 (2024)
21. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
22. He, X., Liu, Q., Qian, S., Wang, X., Hu, T., Cao, K., Yan, K., Zhang, J.: Id-animator: Zero-shot identity-preserving human video generation. arXiv preprint arXiv:2404.15275 (2024)
23. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
24. Hertz, A., Voynov, A., Fruchter, S., Cohen-Or, D.: Style aligned image generation via shared attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4775–4785 (2024)
25. Hleihil, S.: Decartai/Lucy-Edit-ComfyUI. https://github.com/DecartAI/Lucy-Edit-ComfyUI (sep 22 2025), `https://github.com/DecartAI/Lucy-Edit-ComfyUI`
26. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
27. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
28. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
29. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
30. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
31. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
32. Huang, Y., Yuan, Z., Liu, Q., Wang, Q., Wang, X., Zhang, R., Wan, P., Zhang, D., Gai, K.: Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. arXiv preprint arXiv:2501.04698 (2025)
33. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

34. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. arXiv preprint arXiv:2503.07598 (2025)
35. Jones, M., Wang, S.Y., Kumari, N., Bau, D., Zhu, J.Y.: Customizing text-to-image models with a single image pair. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–13 (2024)
36. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
37. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025)
38. Kumari, N., Su, G., Zhang, R., Park, T., Shechtman, E., Zhu, J.Y.: Customizing text-to-image diffusion with object viewpoint control (2024)
39. Kumari, N., Yin, X., Zhu, J.Y., Misra, I., Azadi, S.: Generating multi-image synthetic data for text-to-image customization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16524–16534 (2025)
40. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)
41. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
42. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
43. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36**, 30146–30166 (2023)
44. Liang, F., Ma, H., He, Z., Hou, T., Hou, J., Li, K., Dai, X., Juefei-Xu, F., Azadi, S., Sinha, A., et al.: Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13146–13156 (2025)
45. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
46. Liu, G., Xia, M., Zhang, Y., Chen, H., Xing, J., Wang, Y., Wang, X., Yang, Y., Shan, Y.: Stylecrafter: Enhancing stylized text-to-video generation with style adapter. arXiv preprint arXiv:2312.00330 (2023)
47. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. arXiv preprint arXiv:2502.11079 (2025)
48. Liu, S., Wang, T., Wang, J.H., Liu, Q., Zhang, Z., Lee, J.Y., Li, Y., Yu, B., Lin, Z., Kim, S.Y., et al.: Generative video propagation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17712–17722 (2025)
49. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
50. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
51. Parmar, G., Patashnik, O., Wang, K.C., Ostashev, D., Narasimhan, S., Zhu, J.Y., Cohen-Or, D., Aberman, K.: Object-level visual prompts for compositional image generation. arXiv preprint arXiv:2501.01424 (2025)

52. Patashnik, O., Gal, R., Ostashev, D., Tulyakov, S., Aberman, K., Cohen-Or, D.: Nested attention: Semantic-aware attention values for concept personalization. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12 (2025)
53. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
54. Pexels: Pexels: Free stock photos and videos (2026), `https://www.pexels.com`
55. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
56. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
58. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
59. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
60. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
61. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
62. Rout, L., Chen, Y., Ruiz, N., Kumar, A., Caramanis, C., Shakkottai, S., Chu, W.S.: Rb-modulation: Training-free personalization of diffusion models using stochastic optimal control. arXiv preprint arXiv:2405.17401 (2024)
63. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
64. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
65. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023)
66. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
67. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)
68. Song, K., Zhu, Y., Liu, B., Yan, Q., Elgammal, A., Yang, X.: Moma: Multimodal llm adapter for fast personalized image generation. In: European Conference on Computer Vision. pp. 117–132. Springer (2024)

69. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1921–1930 (2023)
70. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
71. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems **36**, 7594–7611 (2023)
72. Wei, C., Xiong, Z., Ren, W., Du, X., Zhang, G., Chen, W.: Omniedit: Building image editing generalist models through specialist supervision. In: The Thirteenth International Conference on Learning Representations (2024)
73. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15943–15953 (2023)
74. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
75. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. International Journal of Computer Vision **133**(3), 1175–1194 (2025)
76. Xing, P., Wang, H., Sun, Y., Wang, Q., Bai, X., Ai, H., Huang, R., Li, Z.: Csgo: Content-style composition in text-to-image generation. arXiv preprint arXiv:2408.16766 (2024)
77. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023)
78. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
79. Ye, Z., Huang, H., Wang, X., Wan, P., Zhang, D., Luo, W.: Stylemaster: Stylize your video with artistic generation and translation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2630–2640 (2025)
80. Yuan, S., Huang, J., He, X., Ge, Y., Shi, Y., Chen, L., Luo, J., Yuan, L.: Identity-preserving text-to-video generation by frequency decomposition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12978–12988 (2025)
81. Zhang, G., Sohn, K., Hahn, M., Shi, H., Essa, I.: Finestyle: Fine-grained controllable style personalization for text-to-image models. Advances in Neural Information Processing Systems **37**, 52937–52961 (2024)
82. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
83. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al.: Videoprism: A foundational visual encoder for video understanding. arXiv preprint arXiv:2402.13217 (2024)
84. Zhao, M., Wang, R., Bao, F., Li, C., Zhu, J.: Controlvideo: conditional control for one-shot text-driven video editing and beyond. Science China Information Sciences **68**(3), 132107 (2025)
85. Zi, B., Ruan, P., Chen, M., Qi, X., Hao, S., Zhao, S., Huang, Y., Liang, B., Xiao, R., Wong, K.F.: Se\~ norita-2m: A high-quality instruction-based dataset for general video editing by video specialists. arXiv preprint arXiv:2502.06734 (2025)