Video-P2P: Video Editing with Cross-attention Control

Shaoteng Liu[1] Yuechen Zhang[1] Wenbo Li[1] Zhe Lin[3] Jiaya Jia[1,2]
[1]The Chinese University of Hong Kong [2]SmartMore [3]Adobe
https://video-p2p.github.io/

Figure 1: Video-P2P generates new characters while optimally maintaining the pose and environment in videos.

Abstract

This paper presents Video-P2P, a novel framework for real-world video editing with cross-attention control. While attention control has proven effective for image editing with pre-trained image generation models, there are currently no large-scale video generation models publicly available. Video-P2P addresses this limitation by adapting an image generation diffusion model to complete various video editing tasks. Specifically, we propose to first tune a Text-to-Set (T2S) model to complete an approximate inversion and then optimize a shared unconditional embedding to achieve accurate video inversion with a small memory cost. For attention control, we introduce a novel decoupled-guidance strategy, which uses different guidance strategies for the source and target prompts. The optimized unconditional embedding for the source prompt improves reconstruction ability, while an initialized unconditional embedding for the target prompt enhances editability. Incorporating the attention maps of these two branches enables detailed editing. These technical designs enable various text-driven editing applications, including word swap, prompt refinement, and attention re-weighting. Video-P2P works well on real-world videos for generating new characters while optimally preserving their original poses and scenes. It significantly outperforms previous approaches.

1. Introduction

Video creation and editing are key tasks [15, 12, 35, 31, 21]. Text-driven editing becomes one promising pipeline.

Several methods have demonstrated the ability to edit generated or real-world images with text prompts [17, 11, 20]. Till now, it is still challenging to edit only local objects in a video, such as changing a running "dog" into a "cat" without influencing the environment. This paper proposes a pipeline that can edit a video both locally and globally, as shown in Figs. 1 and 5.

Text-driven image editing requires a model capable of generating target content, such as changing the category or property of an object. Diffusion models have demonstrated outstanding generation capabilities in this area [17, 11, 4, 33]. Among these methods, attention control emerges as the most effective pipeline for detailed image editing [11, 20]. In order to edit a real image, this pipeline includes two necessary steps: (1) inverting the image into latent features with a pre-trained diffusion model, and (2) controlling attention maps in the denoising process to edit the corresponding parts of the image. For example, by swapping their attention maps, we can replace a "child" with a "panda".

In this paper, we aim to build an attention control-based pipeline for video editing. Since no large-scale pre-trained video generation models are publicly available, we propose a novel framework to show that a pre-trained image diffusion model can be adapted for detailed video editing.

While a pre-trained image diffusion model can be utilized for video editing by processing frames individually (Image-P2P), it lacks semantic consistency across frames (the 2nd row of Fig. 2). To maintain semantic consistency, we propose using a structure on inversion and attention control for all frames, by transforming the Text-to-image diffusion model (T2I) into a Text-to-set model (T2S). This approach is effective, as illustrated in the 3rd row, where the robotic penguin maintains its consistency across frames.

We adopt the method proposed in [39] to convert a T2I model into a T2S model by altering the convolution kernels and replacing the self-attentions with frame-attentions. This conversion yields a model that generates a set of semantically consistent images. The generation quality will be degraded with the inflation step but it can be recovered after tuning on the original video. Although the tuned T2S model is not an ideal video generation model, it suffices to create an approximate inversion for a video as shown in Fig. 3 (c). It is just an approximation because errors are accumulated in the denoising pass, consistent with conclusions in [20, 36].

To improve the inversion quality, we propose to optimize a shared unconditional embedding for all frames to align the denoising latent features with the diffusion latent features. Our experiments show that shared embedding is the most efficient and effective choice for video inversion. Comparisons are shown in Fig. 3.

As discussed in [11], successful attention control requires a model to have both reconstruction ability and editability. While image inversion has been argued to possess both abilities in [20], we find that video editing presents different challenges. The T2S model, as an inflation model not trained on any videos, is not robust to the perturbations caused by various unconditional embeddings. Although our optimized embedding can achieve reconstruction, changing prompts can destabilize the model and result in a low-quality generation. On the other hand, we find that the approximate inversion with an initialized unconditional embedding is editable but cannot reconstruct well. To address this issue, we propose a decoupled-guidance strategy in attention control, utilizing different guidance strategies for the source and target prompts. Specifically, we use the optimized unconditional embedding for the source prompt and the initialized unconditional embedding for the target prompt. We incorporate the attention maps from these two branches to generate the target video. These two simple designs prove effective and successfully complete video editing. Our contributions can be summarized as:

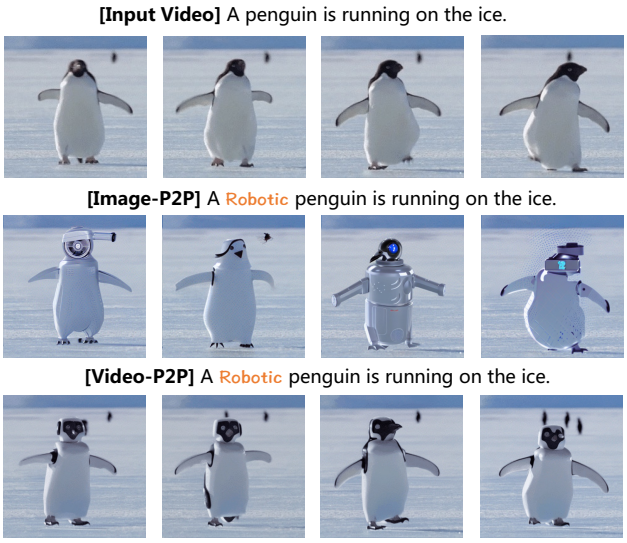

Figure 2: Video-P2P vs Image-P2P. Editing a video frame-by-frame (Image-P2P) cannot guarantee semantic consistency across frames. Video-P2P enables changing the penguin into the same robotic type in every frame.

- We propose the first framework for video editing with attention control. A decoupled-guidance strategy is designed to further improve performance.
- We introduce an efficient and effective video inversion method with shared unconditional embedding optimization to improve video editing substantially.
- We conduct extensive ablation studies and comparisons to show the effectiveness of our video editing framework.

Figure 3: Inversion Comparison. (b) The inflated model cannot generate high-quality results. (c) Tuning the model can create an approximate video inversion. (d) Optimizing a shared unconditional embedding can accurately reconstruct the input video.

2. Related Work

2.1. Text Driven Generation

DALL-E [28] first considers the text-to-image (T2I) generation task as a sequence-to-sequence translation problem, with subsequent research improving generation quality [40, 5, 7]. Denoising Diffusion Probabilistic Models (DDPMs)[13] have recently gained popularity for T2I. GLIDE[22] utilizes classifier-free guidance to improve text conditioning. DALLE-2 [27] leverages CLIP [26] for better text-image alignment. Latent Diffusion Models (LDMs) [29] propose processing in the latent space to enhance training efficiency. In our work, we employ a pre-trained image diffusion model based on LDMs.

Text-to-video (T2V) generation is a nascent research area. GODIVA [38] first introduces VQ-VAE [34] to T2V. CogVideo [16] combines T2V with CogView-2 [5], utilizing pre-trained text-to-image models. Video Diffusion Models (VDM) [15] propose a space-time U-Net for performing diffusion on pixels. Imagen Video [12] successfully generates high-quality videos with cascaded diffusion models and v-prediction parameterization. Phenaki [35] generates videos with time-variable prompts. Make-A-Video [31] combines the appearance generation of T2I models with movement information from video data. While these approaches generate reasonable short videos, they still contain artifacts and do not support real-world video editing. Additionally, most of these approaches are not publicly available at this time.

Several single-video generative models have been proposed. Single-video GANs [1, 10] can generate novel videos with similar objects and motions, while SinFusion [23] uses diffusion models to improve generalization but is limited to simple cases. Tune-A-Video [39] inflates an image diffusion model into a video model and tunes it to reconstruct the input video. It allows for changes in semantic content but with limited temporal consistency. We find that using DDIM inversion results can improve its temporal consistency. However, it cannot avoid altering unrelated regions. We adapt some designs of TAV to do our model initialization.

2.2. Text Driven Editing

Generative models have demonstrated impressive performance in image editing, with approaches ranging from GANs [9, 24, 25, 37] to diffusion models [2, 17]. SDEdit [19] adds noise to an input image and uses the diffusion process to recover an edited version. Prompt-to-Prompt [11] and Plug-and-Play [33] use attention control to minimize changes to unrelated parts, while Null-Text Inversion [20] improves real image editing. InstructPix2Pix [4] enables flexible text-driven editing with user-provided instructions. Textual Inversion [8], DreamBooth [30], and Custom-Diffusion [18] learn special tokens for personalized concepts and generate related images.

Video editing with generative models has seen several advances recently. Text2Live [3] employs CLIP to edit textures in videos but struggles with significant semantic changes. Dreamix [21] uses a pre-trained Imagen Video [12] backbone to perform image-to-video and video-to-video editing, with the ability to change motion as well. Gen-1 [6] trains models jointly on images and videos for tasks such as stylization and customization. While these methods enable modifying video content, they operate like guided generation and tend to modify all regions together when editing an object. Our proposed method allows for local editing with a diffusion model pre-trained on images.

3. Method

Let $\mathcal{V}$ be a real video containing n frames. We adopt the Prompt-to-Prompt setting by introducing a source prompt $\mathcal{P}$ and an edited prompt $\mathcal{P}^*$ which together generate an edited video $\mathcal{V}^*$ containing n frames. The prompts are provided by the user.

To achieve cross-attention control in video editing, we propose Video-P2P, a framework with two key technical designs: (1) optimizing a shared unconditional embedding for video inversion, and (2) using different guidance for the source and edited prompts, and incorporating their attention maps. The framework is illustrated in Fig. 4.

3.1. Preliminary

Latent Diffusion Models (LDMs). LDMs generate an image latent z_0 using a random noise vector z_t and a textual

condition P as inputs. As variants of DDPMs, these models aim to predict artificial noise by minimizing the following objective:

$$\min_{\theta} E_{z_0,\varepsilon\sim N(0,I),t\sim \text{ Uniform }(1,T)} \left\|\varepsilon - \varepsilon_\theta\left(z_t, t, \mathcal{C}\right)\right\|_2^2, \quad (1)$$

where $\mathcal{C} = \psi(\mathcal{P})$ is the embedding of the text prompt, and noise ε is added to z_0 according to step t to obtain z_t. During inference, the model predicts noise $\varepsilon_\theta(\cdot)$ for T steps to generate an image from z_T.

DDIM sampling and inversion. Deterministic DDIM sampling can be used to generate an image from latent features in a small number of denoising steps:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \varepsilon_\theta\left(z_t, t, \mathcal{C}\right). \quad (2)$$

We use an encoder to encode the real image before the diffusion process and a decoder to decode after the denoising process. DDIM sampling can be reversed in a few steps through the equation:

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \varepsilon_\theta\left(z_t, t, \mathcal{C}\right), \quad (3)$$

known as DDIM inversion [32]. This can be used to obtain the corresponding latent features of a real image.

Null-text inversion. To mitigate the amplification effect of text conditioning during image generation, classifier-free guidance is proposed, which performs unconditional prediction [14]:

$$\tilde{\varepsilon}_\theta\left(z_t, t, \mathcal{C}, \varnothing\right) = w \cdot \varepsilon_\theta\left(z_t, t, \mathcal{C}\right) + (1 - w) \cdot \varepsilon_\theta\left(z_t, t, \varnothing\right), \quad (4)$$

where $\varnothing = \psi(\text{""})$ is the embedding of a null text and w is the guidance weight. However, the classifier-free guidance increases errors accumulated in the denoising process, leading to imperfect image reconstruction using the DDIM inversion. [20] proposes to align the diffusion latent trajectory $z_T^*, \ldots, z_0^*$ with the denoising latent trajectory $z_T, \ldots, z_0$ by optimizing a step-wise unconditional embedding $\varnothing_t$:

$$\min_{\varnothing_t} \left\|z_{t-1}^* - z_{t-1}\right\|_2^2. \quad (5)$$

3.2. Video Inversion

We begin by constructing a T2S model that is capable of performing an approximate inversion. Following the VDM baselines [12, 15] and TAV [39], we employ $1\times3\times3$ pattern convolution kernels and temporal attention. Moreover, we replace the self-attentions with frame-attentions, which take the first frames v_0 and the current frame v_i as inputs and

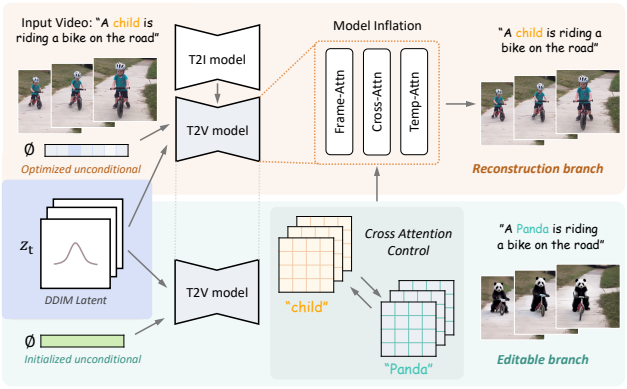

Figure 4: Framework. We optimize one shared unconditional embedding for the reconstruct branch (orange). The initialized unconditional embedding is utilized for the editable branch (green). Their attention maps are incorporated to create the target video.

update features for the frame v_i. The formulation of the frame-attention is as follows:

$$Q = W^Q \mathbf{v}_i, K = W^K \mathbf{v}_0, V = W^V \mathbf{v}_0, \quad (6)$$

where W are the projection matrices in attention. The model processes a video pair-by-pair and computes n times to obtain the prediction for every frame. While the Sparse-causal attention proposed in TAV [39] outperforms frame-attention when generating videos from random noise, we find that the simple design suffices for video inversion since the reversed latent features can capture temporal information. Additionally, frame-attention conserves memory and speeds up the process.

While model inflation can aid in preserving semantic consistency across frames, it adversely impacts the generation quality of the T2I model. This is because the self-attention parameters are utilized to compute frame correlations, which have not been pre-trained. Consequently, the T2S model, generated through inflation, is insufficient for the approximate inversion, as demonstrated in Fig. 2. To address this, we fine-tune the query projection matrices W^Q of the frame- and cross-attentions, as well as additional temporal attention, to perform noise prediction based on the input video following [39]. After this initialization, the T2S model is capable of generating semantically consistent image sets while maintaining the quality of each frame, resulting in successful approximate inversion.

Using the fine-tuned T2S model, we perform video inversion by optimizing a shared unconditional embedding. During inversion, each latent feature z_t contains a channel for the frames with dimension n, where $z_{t,i}$ denotes the latent feature for the i-th frame. We employ DDIM inversion to generate latent features $z_0^*, \ldots, z_T^*$. The unconditional

embedding is defined as follows:

$$\min_{\varnothing_t} \sum_{i=1}^{n} \left\| z^*_{t-1,i} - z_{t-1,i}\left(\bar{z}_{t,i}, \bar{z}_{t,0}, \varnothing_t, \mathcal{C}\right) \right\|_2^2, \text{where} \quad (7)$$

$$\bar{z}_{t-1,i} = z_{t-1,i}\left(\bar{z}_{t,i}, \bar{z}_{t,0}, \varnothing_t, \mathcal{C}\right) \quad (8)$$

is updated at each step. The T2S model's frame-attentions use two latent features to calculate the corresponding feature for the next step. Notice $\varnothing_t$ is shared by all frames ($i = 1, \ldots, n$) which minimizes the memory usage. Besides, using the same unconditional embedding for all frames avoids destabilizing the semantic consistency in attention control.

3.3. Decoupled-guidance Attention Control

To perform attention control on real images, existing works [11, 20] require an inference pipeline with both reconstruction ability and editability. However, achieving such a pipeline for a T2S model is challenging. Video inversion allows us to establish an inference pipeline to reconstruct the original video well. However, the T2S model is not as robust as T2I models due to a lack of pre-training with videos. As a result, its editability is compromised with the optimized unconditional embedding, leading to degraded generation quality when changing prompts. In contrast, we find that using an initialized unconditional embedding makes the model more editable while it cannot reconstruct perfectly. This inspires us to combine the abilities of two inference pipelines. For the source prompt, we use the optimized unconditional embedding in the classifier-free guidance. For the target prompt, we choose the initialized unconditional embedding. We then incorporate attention maps from these two branches to obtain the edited video, where the unchanged parts are influenced by the source branch and the edited parts are influenced by the target branch.

The pseudo algorithm is shown in Alg. 1. We adopt the attention control methods from Image-P2P to Video-P2P. For example, to perform word swap, the $Edit$ function can be represented as:

$$Edit\left(M_t, M^*_t, t\right) := \begin{cases} M^*_t & \text{if } t < \tau \\ M_t & \text{otherwise} \end{cases}, \quad (9)$$

M_t and M^*_t are the cross-attention maps for every frame at every step, and DM is the tuned T2S model. Changing the frame-attentions maps has a small influence on the final results. Attention maps are swapped only for the first τ steps because attentions are formed in the early period. $\overline{M}_{t,w}$ is the average attention map of the word w calculated at step t. It is averaged over steps $T, \ldots, t$ independently for every frame. For the j-th frame, we calculate:

$$\overline{M}_{t,w,j} = \frac{1}{T-t} \sum_{i=t}^{T} \overline{M}_{i,w,j} \quad j = 1, \ldots, n. \quad (10)$$

Algorithm 1: Prompt-to-Prompt video editing

Input: A source prompt $\mathcal{P}$, a target prompt $\mathcal{P}^*$.
Output: Source video V_{src} and edited video V_{dst}.
Latent features from DDIM inversion: z_T;
$z^*_T \leftarrow z_T$;
Initialized unconditional embedding $\varnothing^*$ and optimized unconditional embedding $\varnothing$;
for $t = T, T-1, \ldots, 1$ **do**
  $z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t, \varnothing)$;
  $M^*_t \leftarrow DM(z^*_t, \mathcal{P}^*, t, \varnothing^*)$;
  $\widehat{M}_t \leftarrow Edit(M_t, M^*_t, t)$;
  $z^*_{t-1} \leftarrow DM(z^*_t, \mathcal{P}^*, t, \varnothing^*)\{M^*_t \leftarrow \widehat{M}_t\}$;
  $\alpha \leftarrow B\left(\overline{M}_{t,w}\right) \cup B\left(\overline{M}^*_{t,w^*}\right)$;
  $z^*_{t-1} \leftarrow (1-\alpha) \odot z_{t-1} + \alpha \odot z^*_{t-1}$;
end
Return (z_0, z^*_0)

$B\left(\overline{M}_{t,w}\right)$ represents the binary mask obtained from the attention map. A value is set to 1 when larger than a threshold.

4. Experiments

4.1. Implementation Details

We develop our method based on CompVis Stable Diffusion (v1-5). Similar to TAV [39], we fix the image autoencoder and sample 8 or 24 frames at the resolution of 512×512 from a video. To initialize the model, we fine-tune the T2S model for 500 steps to reconstruct the original video. During attention control, we set the cross-attention replacing ratio to 0.4 and the attention threshold to 0.3. For prompt refinement, we set the refinement ratio to 0.4. These parameters can be adjusted to control the editing fidelity for different examples. All 8-frame experiments are conducted on a single V100 GPU, with 5 minutes for initialization (tuning), 6 minutes for inversion, and 1 minute for inference.

4.2. Applications

Our Video-P2P method can be utilized for a range of editing applications, including prompt refinement, attention re-weighting, and word swapping, similar to the capabilities of image-P2P. Video-P2P is able to maintain semantic consistency across different frames and preserve the temporal coherence of the original video during the editing process. More examples can be found in the appendix.

Word swap. Video-P2P enables the replacement of entities based on word swapping while maintaining the coherence of unrelated regions. As illustrated in Fig. 5, Video-P2P seamlessly replaces the man on the motorbike with Spider-Man while minimizing the changes to the motorbike's appearance (the 4th row). The generated Spider-Man exhibits

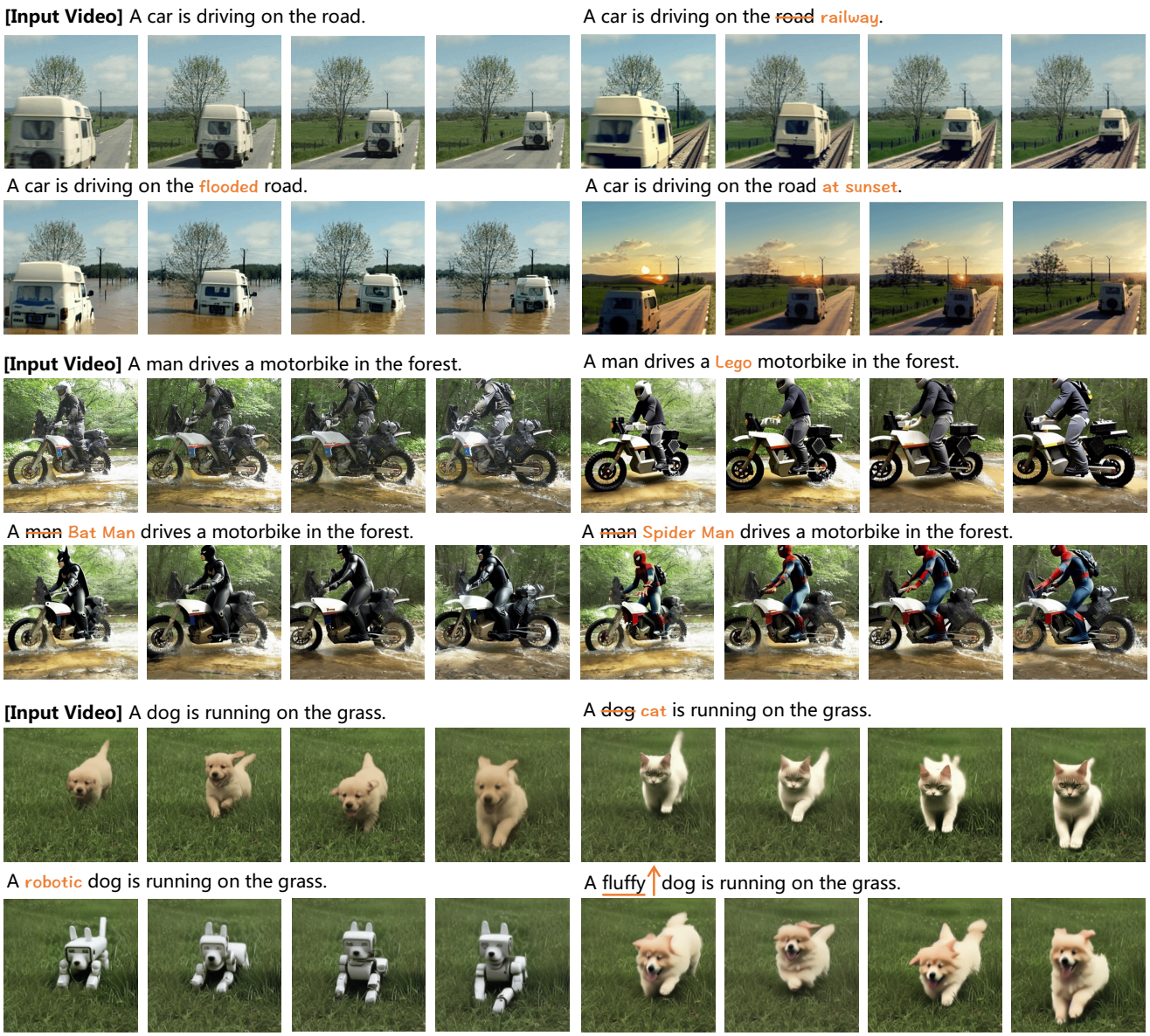

Figure 5: Videos edited by Video-P2P with text prompts. Video-P2P can do both word swaps and prompt refinement.

a consistent appearance across frames, and the background remains unchanged. Furthermore, we can replace a dog with a cat while preserving its gesture and the surrounding grass (the 5th row).

Prompt refinement. Video-P2P is able to do prompt refinement, such as modifying object properties. For example, we can transform the running dog into a robotic one (the 6th row in Fig. 5), and convert a motorbike into a Lego toy with the same motion (the 3rd row). Notice the grass and sky are almost not influenced. Additionally, Video-P2P can perform global editing like changing the weather to sunset or flooding the road with water (2nd row). Style transfer can also be accomplished by Video-P2P, as exemplified by transforming the video into a watercolor painting.

Attention re–weighting. Similar to Image-P2P, Video-P2P also enables attention re-weighting. By adjusting the cross-attention of specific words, we can manipulate the extent of the corresponding generation. For instance, we can regulate how fluffy a dog is in the video (the 6th row of Fig. 5).

4.3. Comparison

Comparison with Tune-A-Video. Both TAV+DDIM [39] and our Video-P2P allow for video editing with text prompts. However, TAV+DDIM cannot avoid altering the entire video content when editing specific objects, while Video-P2P can edit a local area and minimize the influence

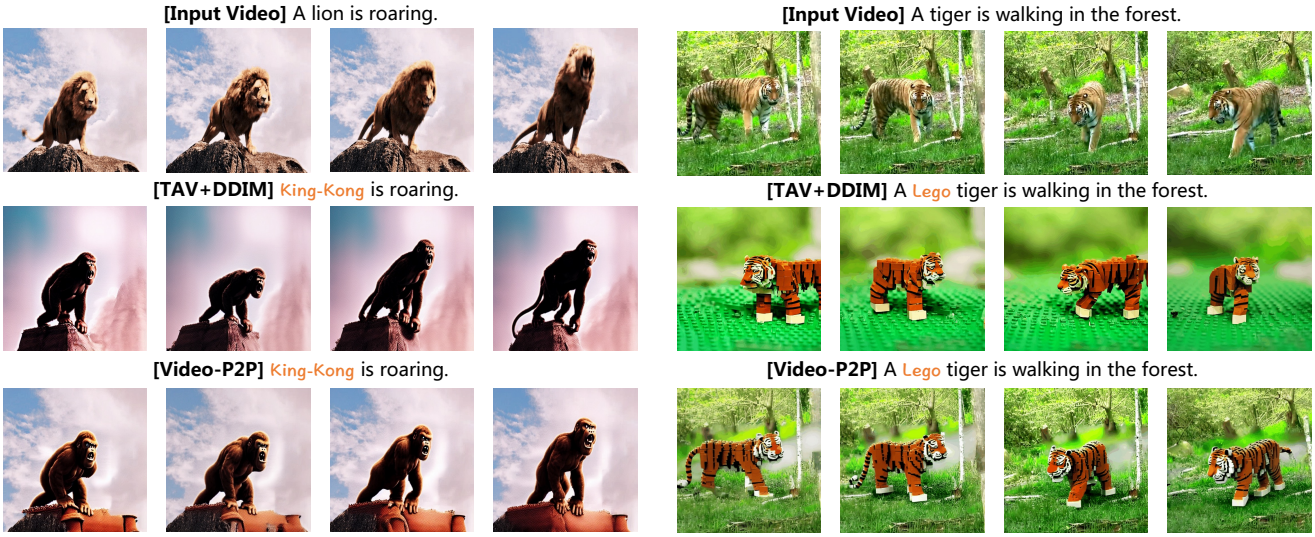

Figure 6: Video-P2P v.s. Tune-A-Video (TAV). Video-P2P offers the ability to edit content locally, while TAV+DDIM cannot avoid influencing unrelated regions.

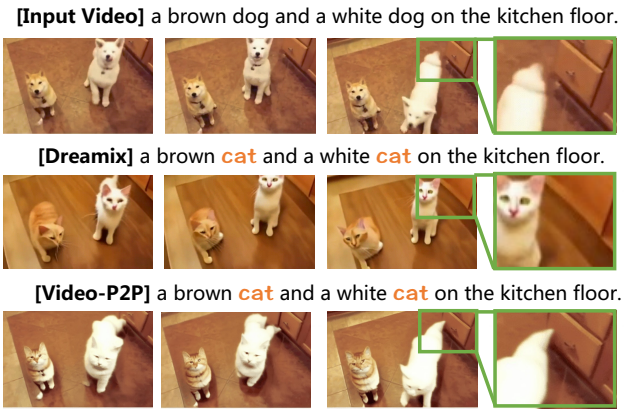

Figure 7: Video-P2P v.s. Dreamix. Both methods can change the objects to new categories. Only Video-P2P can preserve the details in the background.

on other regions. Fig. 6 (Left) demonstrates that Video-P2P preserves the complex shape of the cloud when replacing a lion with King Kong, whereas TAV+DDIM can only maintain the color tone of the sky in this case.

Although our model initialization is similar to TAV, Video-P2P can still generate temporal-consistent results where TAV+DDIM fails. As demonstrated in Fig. 6 (Right), TAV struggles to generate a temporally consistent sequence in the second row, even when the inputs are features from DDIM inversion. In contrast, our method can produce better structure-preserved results, as shown in the third row.

Comparison with Dreamix. In contrast to Dreamix [21], which uses a pre-trained video diffusion model that is not publicly available, our method yields superior results for subject replacement. Although our method cannot perform video motion editing due to the lack of temporal priors, we outperform Dreamix in preserving details and motion consistency. As Dreamix is not open-sourced, we conducted our evaluation on its released demo. As demonstrated in Fig. 7, both methods can transform two dogs into two cats, but our method preserves the details of the drawer in the background (the 3rd row). Furthermore, Dreamix may affect the time sequence to some extent, as the generated cat moves more slowly than the original dog in the video. In contrast, our method completely preserves the motion of the original video.

Quantitative results. We evaluate our proposed Video-P2P on 10 YouTube videos and report four metrics for quantitative analysis. The CLIP Score measures the textual similarity between the text prompt and video, while Masked PSNR and LPIPS [41] evaluate the quality of structure preservation. We also proposed a novel metric, Object Semantic Variance (OSV), to measure semantic consistency across frames. For detailed explanations of these metrics, please refer to the appendix. Our results, as shown in Table 1, demonstrate that Video-P2P performs well on all metrics. Compared to TAV+DDIM, Video-P2P achieves higher Masked PSNR and lower LPIPS, indicating better preservation of unchanged regions. Compared to the other two methods, Video-P2P has a much lower OSV, indicating its superior ability to maintain semantic consistency across frames. Moreover, in Tab. 3, we report the user study results, where Video-P2P ranks first on average and has a high preference rate compared to other methods.

4.4. Ablation Study

Model initialization. While the inflated image diffusion model can generate semantically consistent images, the T2S model's generation ability is compromised during inflation,

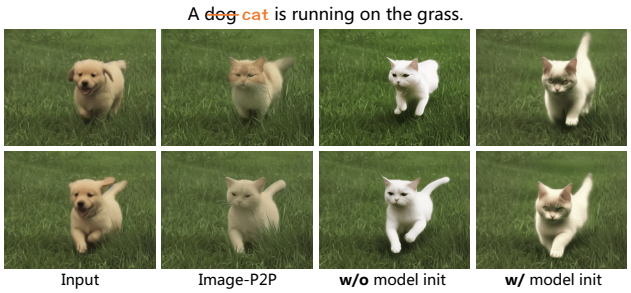

Figure 8: Ablation on model initialization. Though Video-P2P can still generate consistent content without model initialization, the generation quality is degraded. With the model initialization, the generation quality is recovered.

Figure 9: Ablation on decoupled-guidance attention control. Using decoupled-guidance improves the generation quality.

making it inadequate for video inversion even with an optimized unconditional embedding. As seen in Fig.8 (the 3rd column), directly using the inflated T2S model produces unrealistic results with an inaccurate background. To mitigate this, we initialize the T2S mode by fine-tuning the given video. This is evident in Fig.8 (4th column), where the cat's appearance improves, and the grass reconstruction becomes more accurate.

Shared unconditional embedding. Table 2 presents the quantitative results for video inversion. We observe that optimizing a shared unconditional embedding can significantly improve the PSNR compared to TAV+DDIM. However, using multiple unconditional embeddings for each frame only increases the PSNR by 0.2 but results in a higher parameters usage (n times). Besides, we find that using multiple unconditional embeddings leads to a lower Masked PSNR of 20.51 after attention control compared to the shared unconditional embedding. Thus, we conclude that shared unconditional embedding is the most effective and efficient method for video inversion.

Decoupled-guidance attention control. To obtain the latent features of the input video, we optimize an unconditional embedding using the source prompt. It is important to note that this embedding is only suitable for the source

	CLIP ↑	M.PSNR ↑	LPIPS ↓	OSV ↓
TAV+DDIM	0.3322	17.32	0.4625	55.12
Image-P2P	0.3269	22.99	0.3065	76.50
Ours (w/o DG)	0.3224	18.96	0.3864	68.76
Ours	0.3361	20.54	0.3297	47.57

Table 1: Quantitative evaluation. We evaluate editing textual similarity (CLIP Score), region preservation (Masked PSNR, LPIPS), and Object Semantic Variance (OSV) for semantic consistency. DG refers to Decoupled-Guidance.

	VQVAE	TAV +DDIM	Multi-uncond	**Shared-uncond**
PSNR(dB) ↑	24.73	15.43	22.97	22.75
#Param. ↓	/	0.13M	22.68M	2.94M

Table 2: Reconstruction quality on video inversion. A shared unconditional embedding can reconstruct a high-quality video with a small size.

	Image-P2P	TAV	TAV+DDIM	**Video-P2P**
Structure	2.67	3.33	2.61	1.39
Preserving	13.59%	6.52%	10.87%	69.02%
Text	3.40	2.78	2.28	1.54
Alignment	3.80%	14.13%	19.57%	62.50%
Realism &	3.38	2.98	2.21	1.43
Quality	4.35%	7.61%	19.02%	69.02%

Table 3: User study result of average ranking ↓ and preference rate ↑.

prompt during the prompt-to-prompt process. Using the optimized embedding for the target prompt may negatively impact the quality of the generated results, as shown in Fig. 9 (1st row). Instead, we utilize the initialized unconditional embedding for the target prompt and incorporate attention maps from two branches. The decoupled-guidance attention control approach significantly improves the editing quality, as shown in Fig.9 (the 2nd row). Quantitative ablations can be found in Tab. 1 (the 3rd row and 4th row).

5. Conclusion

Our proposed approach, Video-P2P, provides a simple yet effective solution for video editing with cross-attention control. By leveraging a pre-trained image diffusion model, we demonstrate that editing a video locally and globally is possible. Specifically, we optimize a shared unconditional embedding based on a well-initialized T2S model for video inversion. We also propose using different unconditional embeddings for source and target prompts, and integrating attention maps from two branches for improved attention control. These techniques enable Video-P2P to perform various applications, such as word swap, prompt refinement, and attention re-weighting. In future work, we will enhance its capability to handle more complex editing tasks like injecting extra objects.

References

[1] Rajat Arora and Yong Jae Lee. Singan-gif: Learning a generative video model from a single gif. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1310–1319, 2021. 3

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 3

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723, 2022. 3

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 2, 3

[5] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 3

[6] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 3

[7] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106, 2022. 3

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[9] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, pages 1–13, 2022. 3

[10] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical patch vae-gan: Generating diverse videos from a single sample. *Advances in Neural Information Processing Systems*, pages 16761–16772, 2020. 3

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 3, 5

[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 3, 4

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 3

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 1, 3, 4

[16] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[17] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2, 3

[18] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 3

[19] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[20] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2, 3, 4, 5

[21] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 1, 3, 7

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[23] Yaniv Nikankin, Niv Haim, and Michal Irani. Sinfusion: Training diffusion models on a single image or video. *arXiv preprint arXiv:2211.11743*, 2022. 3

[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 3

[25] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. 3

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.

Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831, 2021. 3

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 3

[31] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1, 3

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[33] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 2, 3

[34] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 2017. 3

[35] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 1, 3

[36] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. *arXiv preprint arXiv:2211.12446*, 2022. 2

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3

[38] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 3

[39] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022. 2, 3, 4, 5, 6

[40] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 3

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7