INFUSION: Inject and Attention Fusion for Multi Concept Zero-Shot Text-based Video Editing

Anant Khandelwal
Glance AI
anant.iitd.2085@gmail.com

Source Prompt: A silver jeep driving down a curvy road in the countryside.

Zero Shot Multi Concept Editing with Stable Diffusion v1.5 **Silver Jeep** → Porsche car, **countryside** → Landmark of autumn

Zero Shot Multi Concept Editing with Stable Diffusion v1.5 **Silver Jeep** → Porsche car, **countryside** → snowy winter

Abstract

Large text-to-image diffusion models have achieved remarkable success in generating diverse, high-quality images. Additionally, these models have been successfully leveraged to edit input images by just changing the text prompt. But when these models are applied to videos, the main challenge is to ensure temporal consistency and coherence across frames. In this paper, we propose InFusion, a framework for zero-shot text-based video editing leveraging large pre-trained image diffusion models. Our framework specifically supports editing of multiple concepts with pixel-level control over diverse concepts mentioned in the editing prompt. Specifically, we inject the difference in features obtained with source and edit prompts from U-Net residual blocks of decoder layers. When these are combined with injected attention features, it becomes feasible to query the source contents and scale edited concepts along with the injection of unedited parts. The editing is further controlled in a fine-grained manner with mask extraction and attention fusion, which cut the edited part from the source and paste it into the denoising pipeline for the editing prompt. Our framework is a low-cost alternative to one-shot tuned models for editing since it does not require training. We demonstrated complex concept editing with a generalised image model (Stable Diffusion v1.5) using LoRA. Adaptation is compatible with all the existing image diffusion techniques. Extensive experimental results demonstrate the effectiveness of existing methods in rendering high-quality and temporally consistent videos.

1. Introduction

With the rise in the creation and consumption of video content on social media platforms, there is a need for generalised video creation and editing tools. Despite the recent success of text-to-image diffusion models, their applicability to video is limited since per-frame editing does not produce consistent editing across all the frames. To overcome this limitation, recent research introduced three types of text-to-video diffusion: a) first solution is to train the model on large-scale video data [8] which require lot of computing resources b) second solution is to fine-tune the image models on single video [31] c) third solution is the zero-shot method [11, 19], which requires no training, is compatible with pre-trained image diffusion models, and requires fewer computing resources. In this paper, we employ the zero-shot strategy for text-based video editing. However, the challenges associated with zero-shot methods are: 1) Temporal Consistency: Cross-Frame Continuity 2) Zero-Shot: no training or fine-tuning required 3) Flexible: compatible with off-the-shelf, pre-trained image models. In this paper, we demonstrated the use of a large-scale pre-trained text-to-image model (i.e., Stable diffusion v1.5 [22]), which contains almost all the concepts, hence can be used for any customised generation as opposed to the zero-shot method Fatezero [19], which requires a one-shot tuned model for customised generation.

In this paper, we introduce a novel zero-shot framework for text guided video editing with fine grained control over multiple concepts. Our framework, INFUSION, consists of two parts INJECT and ATTENTION FUSION. In the first part, we inject features from residual block in decoder layers and attention features (obtained from source prompt P_s) into the denoising pipeline for editing prompt P_e. This injection step highlights the target concepts since we injected the difference between residual block features for (P_s, P_e) and combined them with attention injection (keys and values) to query the source contents, keeping the unedited concepts as they are and scaling up the edit concepts in the edit pipeline while scaling down the removed concepts from the source pipeline. In the second part, we fuse the attention for edited and unedited concepts using the mask extraction obtained from cross-attention maps for P_e and P_s, respectively. The fused attention preserves the source content with editing concepts. Additionally, we mix the cross-attention from the source and edited prompts to contain the unedited and edited concepts, respectively. To summarize, our main contributions are as follows:

- A novel zero-shot framework capable of editing multiple concepts with finer details with a single editing pipeline. It achieves the best temporal consistency and generates coherent edited videos, with no training involved either for the generalised image diffusion model or for edited video generation.
- INJECT for fine-grained control over editing concepts and ATTENTION FUSION to cut the edited part and paste the unedited part from source attention.
- Experimental results demonstrate the flexible structure, shape, colour, and style of editing with a temporally coherent generation of edited videos.

2. Related Work

Large-scale zero-shot methods for text-based image editing triggered interest in videos as well. Recent developments introduced video editing methods, namely, Tune-A-Video [31] is a one-shot method that inflates an image diffusion model into a video model with cross-attention and generates edited video by fine-tuning on a single video. Other methods based on the same idea are Edit-A-Video [25], VideoP2P [14] and vid2vid-zero [30] which uses Null-text inversion [15]for preserving unedited regions. However, all these methods require fine-tuning of the pre-trained model over the input video. Following these zero-shot methods are introduced, namely, FateZero [19] proposed attention blending using features before and after editing, Text2Video-Zero [11] denoise the latent directly to motions, Pix2Video [4] matches the current frame with the previous frame in latent space. All the mentioned zero-shot methods largely rely on manipulation with cross-attention maps for early-step latent fusion to improve temporal consistency. However, as we demonstrate, these methods are effective in editing high-level styles and shapes but less effective in manipulating concepts at fine-grained levels. Our method does the editing at finer levels using feature injection, which acts at pixel level. Over and above, we apply attention feature injection and fusion to control over the concepts mentioned for editing.

3. Preliminary

Latent Diffusion Models: Diffusion models [22, 9, 17, 26] are probabilistic generative models that can generate the desired image from an initialised Gaussian noise image $x_T \sim \mathcal{N}(0, \mathbf{I})$ by progressively removing the noise at step ranging from T to 0. In general, the foundation of diffusion models is based on two complementary random processes i.e. *forward* and *backward*. During *forward process* or *inversion* the noise is added at each step from 0 to T to clean image x_0 defined as:

$$x_t = \sqrt{\alpha_t} \cdot x_0 + \sqrt{1 - \alpha_t} \cdot z \tag{1}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$ and α_t are the noise schedule.

The *backward process* or *reconstruction* is aimed at progressively denoising the image x_T, where at each step t the cleaner version of image is obtained than the previous step $t+1$, and finally to cleaned image at 0. This is achieved by a neural network $\epsilon_\theta(x_t, t)$, which predicts the added noise

z at each step. Once trained, this is applied at each backward step which consists of applying ϵ_θ to the current x_t, and adding a Gaussian noise perturbation to obtain a cleaner x_{t-1}, defined as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t), \tag{2}$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \epsilon\frac{\sqrt{1-\alpha_t}}{\sqrt{1-\bar{\alpha}_t}}), \tag{3}$$

where $\bar{\alpha}_t = \prod_i^t \alpha_t$, and ϵ is the predicted noise. Neural network ϵ_θ is trained using the mean squared error given as:

$$\mathcal{L} = \mathbb{E}_{x_0,\epsilon,t}(||\epsilon - \epsilon_\theta(x_t, t)||) \tag{4}$$

Diffusion models are evolving very fast and have been integrated and trained to generate images conditioned on multiple types of guiding signals, denoted as y in $\epsilon_\theta(x_t, y, t)$ i.e. another image [24], text [12, 16, 21, 22] or class label [10]. In this work, we leveraged the pre-trained text-conditioned Latent Diffusion Model (LDM), a.k.a. Stable Diffusion [22], which performs the diffusion-denoising process in the latent space of the pre-trained image auto-encoder network. The structure of the denoising backbone ϵ_θ is realized as a time-conditional U-Net [23] conditioned on the guiding text prompt P.

Self-Attention and Cross-Attention: Layers of denoising U-Net consists of a residual block [6], a self-attention block and a cross-attention block [29]. At the denoising step t, the residual block convolves features from previous layer ϕ_t^{l-1} to produce the intermediate features f_t^l at the layer l. In self-attention block these intermediate features are projected to produce the queries q_t^l, keys k_t^l and values v_t^l. The output feature of self-attention is then given as:

$$\hat{f}_t^l = A_t^l v_t^l, \text{ where } A = Softmax(q_t^l {k_t^l}^{\mathbf{T}}) \tag{5}$$

Finally, the textual prompt P features are projected into keys and values, which are queried by self-attended spatial features, which when plugged into the attention equation 5 will compute the features at the output of the cross-attention block. These attention maps in the stable diffusion collectively contain the rich information of structure, shape, and layout present in the spatial features obtained from residual blocks. Cross-attention maps the spatial pixels to the input text prompt and allows the editing [7] of multiple granular objects that are present in the source video. Meanwhile, the features in self-attention layers are employed in a plug-and-play manner [28] to retain the structure/ layout/ shape of un-edited objects and facilitate style editing over them with the edited prompt. Collectively, in this work, we leveraged the combination of cross-attention and self-attention maps to perform consistent video synthesis with delicate handling of multi-concept editing in a zero-shot manner that can preserve and retain the layout and structure of edited and un-edited parts in a prompt, respectively.

4. InFusion

In this section we present *InFusion*, a framework designed to do zero shot text based video editing of multiple concepts and ensuring temporal consistency between frames of the edited video. Formally, given the source input video $X_0 = \{x_0\}_{i=1}^N$ with N frames, source prompt P_s and the target prompt P_e, the goal of text driven video editing is to generate a video $Y_0 = \{y_0\}_{i=1}^N$ which aligns with prompt P_e, faithfully preserves the unedited content of source video X_0 and maintains the temporal consistency between frames. Our framework is built upon Stable Diffusion v1.5 [22] a pre-trained and fixed text-to-image LDM model denoted by $\epsilon_\theta(x_t, P, t)$ where P is the given prompt. This model is based on the U-Net architecture with T time-steps denoising as shown in Figure 1 and discussed in Section 3. However, this model can generate the frames as per the given prompt but to ensure temporal consistency between frames and retaining the unedited contents of source video we made several modifications to the pipeline.

Our key finding is that the fine grained control over the generated structure is achieved by highlighting each concept using the a) edit directions obtained from difference of spatial features from source and edit prompts b) accurate mask extraction from source and edited cross-attention maps for fine grained control over the structure of edited shape c) retain the unedited structure by combining cross-attention maps from source and edit prompts for source and edited parts respectively.

4.1. INJECT

Spatial Features: Spatial features in text-to-image generation methods govern the basic part of specifying the structure/shape/pose/scene layout. Even if the prompt is descriptive, like "*a Porsche car driving down a curvy road in the countryside*" or "*a cat jumping over the bed*" the model can generate different images under different initial noise x_T. We hypothesise that in text-based editing, the structure/pose can be controlled in a fine-grained manner using the spatial features, and this hypothesis is motivated by the analysis in [2, 28], which demonstrated the semantic segments obtained from spatial features. To further investigate this fact, we did a PCA analysis, as shown in Figure 2. Specifically, for each input image, we extract the features f_t^l from each layer in decoder of ϵ_θ at each time-step and compute the first three principal components as displayed in Figure 2 for layers 4,7 and 11. As seen, in the coarsest layer (layer 4), a crude blob of Jeep structure is visible, but in layers 7 and 11, the Jeep structure is clearly visible. Interestingly, the colour of the similar object (irrespective of its pose) is same across all the frames at each layer. In text-based video editing, we have to retain the source layout, and hence we choose to inject the source features while editing with the prompt P_e, but while

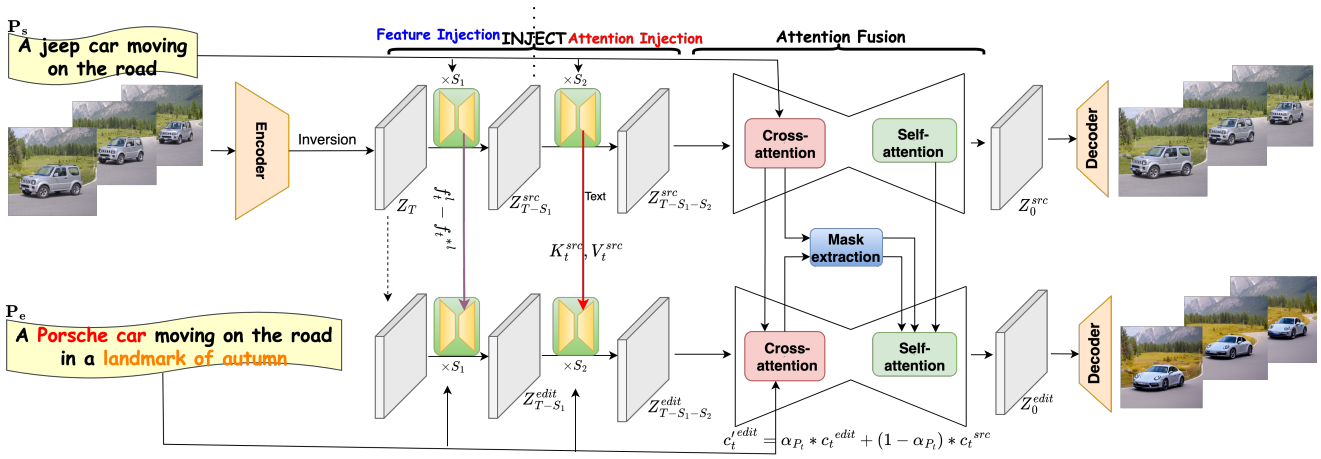


Figure 1: INFUSION: Leveraging a pre-trained text-to-image model for video editing ensures temporal consistency and editing accuracy with Inject and Attention Fusion. The denoising pipeline for source prompt P_S generates the decoder latent from U-Net and attention features from source video, which are injected into the denoising pipeline (initialised with inverted source latent z_T) for edit prompt P_e.

injecting, we have to edit the structure of some objects, so we choose to inject the source features in coarse layers only, since features at higher layers gradually capture more fine-grained information based on the features at coarse layers, and since these features eventually contribute to the error predicted by the U-Net, the tendency will be more towards decreasing the error at finer levels.

Feature Injection and Edit Direction: We now discuss the translation of the given source (x_0, P_s) to edited video y_0 from the edited prompt P_e. First, the source video is inverted using DDIM[26] to noise denoted as z_T. Given the target prompt P_e, the generation of edited video y_0 is performed using the same initial noise z_T as shown in Figure 1. At each step t of the backward process from initial noise z_T for source prompt, the guidance features $\{f_t^l\}$ are collected at each layer from the denoising step $z_{t-1} = \epsilon_\theta(x_t, P_s, t)$. We then inject these source guidance features $\{f_t^l\}$ during the denoising steps of y_t from the target prompt P_e. Specifically, we replace the resulting features $\{f_t^{*l}\}$ given as follows:

$$z^*_{t-1} = \epsilon_\theta(y_t, P_e, t; \{f_t^l - f_t^{*l}\}) \text{ where } y_T = z_T \quad (6)$$

The edited prompt "*a Porsche car driving down a curvy road in a landmark of autumn*" contains the following edited concepts: a) "***silver jeep*** $\rightarrow$ *Porsche car*" b) "***countryside*** $\rightarrow$ *landmark of autumn*". As shown in Figure 2 the spatial features $\{f_t^l - f_t^{*l}\}$ at *layer 4*, the jeep structure which was visible in f_t^l for source prompt P_s is now not clearly visible (looks moving towards car structure in some frames), and the colour of the "*countryside*" is also changed, depicting that it is moving from source concepts to edited concepts. Hence, we instead injected the $\{f_t^l - f_t^{*l}\}$ features since we want to move from source concepts to edited concepts as mentioned in a) and b) rather than retaining them. However, to reflect the complete structure change for edited concepts, feature injection is not enough since it can only give the edit directions (after a few steps S_1 as shown in Figure 1), these are leveraged in controlling the self-attention to cut the edited concepts from the source structure and paste the remaining part without any modification.

Self-Attention Control: Figure 3a depicts the mechanism of self-attention control, where the keys K_t^{sl} and values ${V_t}^{sl}$ from source (collected during denoising step of x_t) are injected during denoising step of y_t. Specifically, as shown in Figure 3a the queries Q obtained from the injected spatial features $\{f_t^l - f_t^{*l}\}$ downscale the affinities for the edited concepts (a and b as discussed in feature injection) due to the structure changed for those concepts in feature injection and hence less similarity for those concepts with keys K_t^{sl}, less probability for the edited concepts, which scale down the edited concepts in ${V_t}^{sl}$. This ensures the self-attention needed to control the propagation of edited parts. However, the cross-attention maps, which correlate to the target prompt using keys and values, make the edit concepts slowly integrate into the source layout. Once the source layout with edited concepts is brewed in the diffusion process (after S_2 steps as shown in Figure 1) we perform the mask-guided mixing of self-attention and cross-attention from source and target prompt. The INJECT operation is defined as follows:

$$z^*_{t-1} := \begin{cases} \epsilon_\theta(y_t, P_e, t; \{f_t^l - f_t^{*l}\}), t \in [0, S_1), l < L \\ \epsilon_\theta(y_t, P_e, t; \{K_t^{sl}, {V_t}^{sl}\}), t \in [S_1, S_2], \forall l \end{cases} \quad (7)$$

4.2. ATTENTION FUSION

Cut and Paste Self-Attention: We observed that synthesised videos using the INJECT operation faithfully generate the source layout with noise in the edited concepts like overlapping parts of "*jeep*" and "*Porsche car*". Hence, we propose the use of mask-guided editing of self-attention maps for faithful reconstruction of edited concepts without

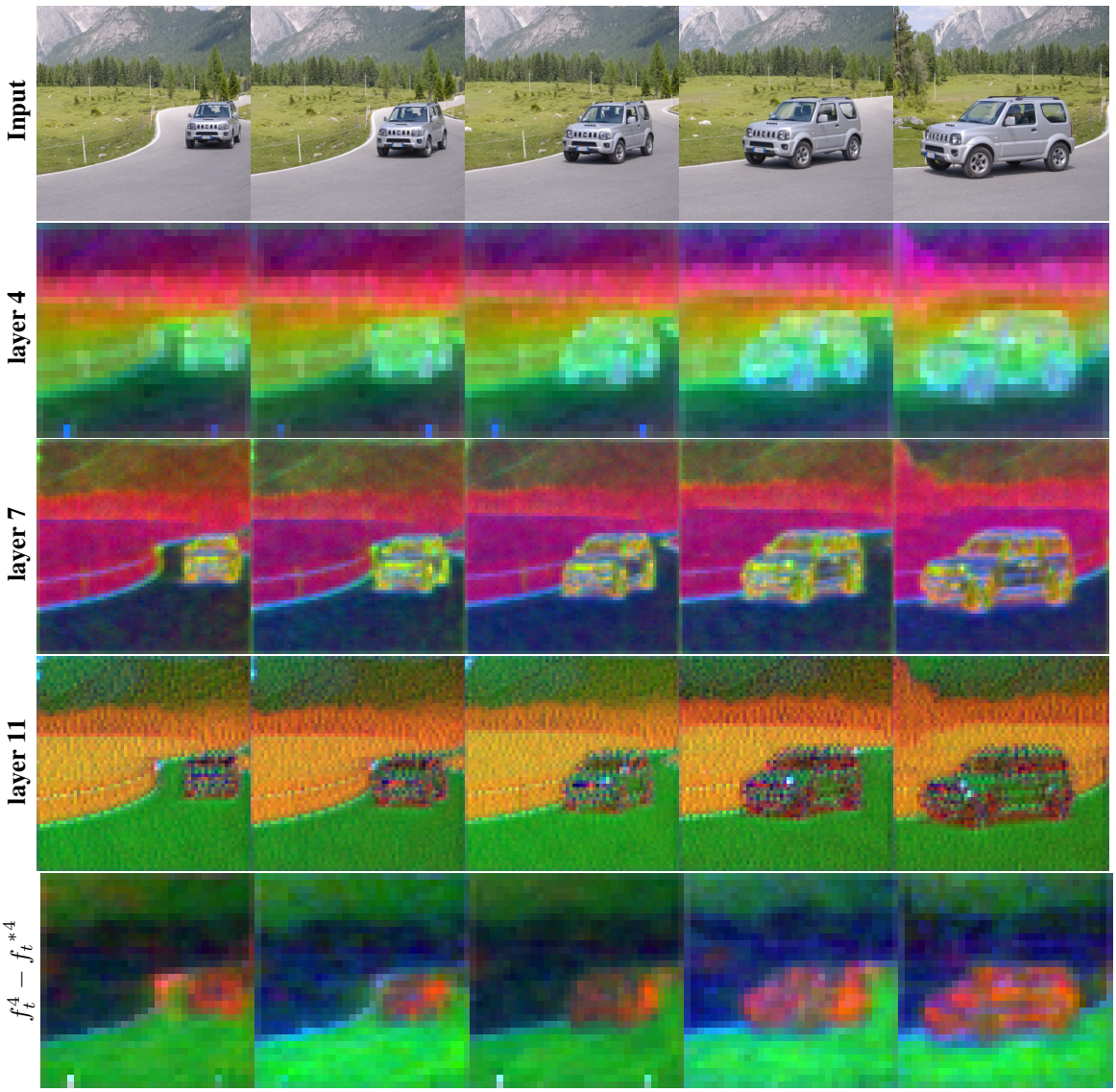


Figure 2: Visualising Top-3 principal components of diffusion features (spatial features) obtained from the decoder of U-Net at different layers.

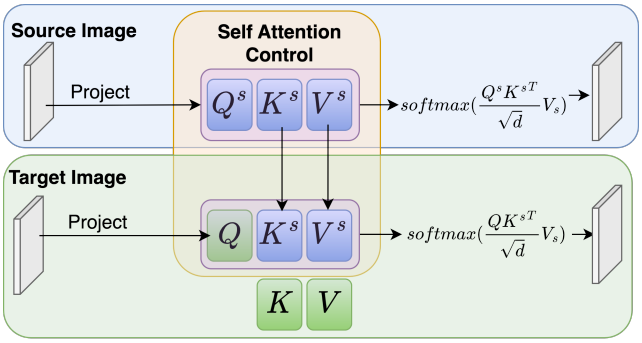
any overlapping parts from the source for those edited concepts. Inspired by the previous works [7, 27, 3], it is revealed that cross-attention maps correlate to the target prompt and hence can be used to inject the edited concepts from the attention maps obtained from the edit prompt while keeping the source layout. Specifically, at step t, we store all the self-attention and cross-attention maps from source prompt P_s and edit prompt P_e during denoising steps using fixed backbone U-Net with INJECT in place for the edit prompt P_e. Then we average the source and edit cross-attention maps for edited words in P_e across all the heads and layers with spatial resolution 16×16, the resulting maps are denoted as $A_c^{t^{src}} \in R^{16\times16\times N}$ (N is the number of tokens edited) and similarly in target attention for edited words denoted as $A_c^{t^{edit}} \in R^{16\times16\times N}$. We then calculate the masks (shown in Fig. 3b) by thresholding the max pooled maps of both $A_c^{t^{src}}$ and $A_c^{t^{edit}}$ denoted as M_s and M_e respectively. This captures only the foreground edited objects in the binary mask. The resulting mask-guided self-attention is given as:

$$s_t^{fused} = M_e * s_t^{edit} + (1 - M_s) * s_t^{src} \quad (8)$$

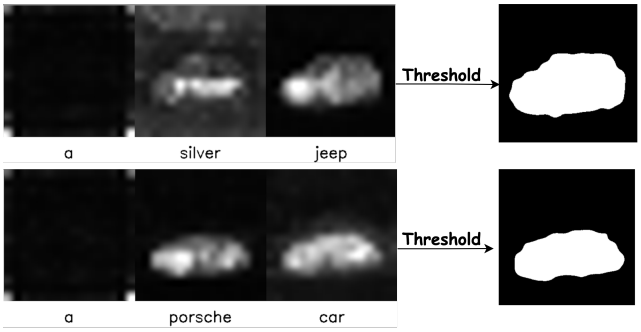
$$s_t^{fused} = fill(equal(s_t^{fused}, 0), s_t^{edit}) \quad (9)$$

which denotes that we take the foreground edited objects from target attention maps only, which is "*cut*" and "*paste*" the remaining background from source using $1 - M_s$.

Cross-Attention Fusion: Similarly, the final fused cross-attention is obtained by taking the target cross-attention maps c_t^{edit} for all heads and all layers for the edited words and taking the source attention c_t^{src} for unedited words.



(a) Self-Attention Control



(b) Mask Extraction from cross-attention maps

Figure 3: a) Self-attention control to query contents from source image in decoder part of U-Net b) Mask extraction strategy

Mathematically, the fused cross-attention is given as:

$$c_t^{fused} = \alpha_w * c_t^{edit} + (1 - \alpha_w) * c_t^{src} \tag{10}$$

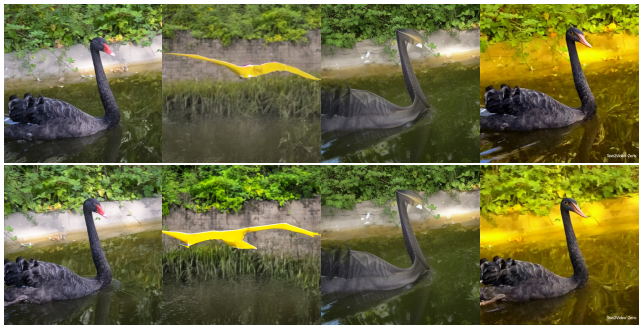
where α_w is the $1/0$ array indicating 1's where the word index is edited and 0 for source words in the edit prompt P_e as compared to source prompt P_s.

4.3. Spatio-Temporal Attention

Inspired by the previous works [31, 19, 33] we also leveraged the spatio-temporal attention for consistent video synthesis for the edit prompt. Since it has been observed that spatial features are the basic foundation of structure in the synthesised video, we initialised the weights of temporal attention with weights of spatial self-attention. Specifically, we integrated the key frame attention into the spatial self-attention to align all the frames with the key frame. Formally, let z^i and z^k denote the embedding of the i-th frame and key frame, respectively. The modified spatial attention is given as:

$$Q = W^Q z^i, K = W^K[z^i; z^k], V = W^V[z^i; z^k] \tag{11}$$

where [;] denotes concatenation, W^Q, W^K, W^V are the projection matrices of the pre-trained model. Among the possible key-frame choices: a) $k = \text{round}(\frac{N_F}{2}$ where N_F is the total number of frames) b) $k = i - 1$ c) $k = i + 1$. We find that there are no significant differences in using any of the above choices, and hence, for ease, we choose to take k=i-1 as the key frame. The resulting spatio-temporal attention map is represented as $s_t \in R^{hw \times 2hw}$ where 2 denotes the spatio-temporal correspondence considered in calculating attention at a given time step. Overall, this concludes the end-to-end zero shot editing with the general-purpose diffusion model, which requires no fine tuning for editing the concepts that are already present in the trained model.



(a) Input **(b)** Ours **(c)** FateZero **(d)** T2V-Zero

Figure 4: Qualitative comparison of our method with FateZero and Text2Video-Zero (T2V-Zero). Best viewed with zoom-in

5. Experimental Results

5.1. Implementation Details

All the experiments are conducted on one NVIDIA Tesla A100 40GB GPU. For zero-shot text-based video editing, we use the trained Stable Diffusion 1.5 [22] as the base text-to-image model, which has been converted to a video editing model by incorporating spatio-temporal attention along with Inject and Attention Fusion as shown in Figure 1. Throughout all experiments, the DDIM sampler was used with $T = 50$ steps and 7.5 as the classifier guidance for the video editing pipeline. Out of 50 steps, the total number of steps for Inject operation is 12, of which 6 are for feature injection and the rest 6 are for self-attention injection, denoted as S_1 and S_2 in Figure 1 respectively. We explained above the choice of feature injection for coarse layers only, hence L = 6 in our case. The mask threshold is set to 0.3 for the case of editing "*Porsche car*", but it is subject to change from case to case. Rest of the steps are consumed by Attention Fusion operation. Following existing works [19, 33, 1, 5], we evaluated our method on videos from DAVIS[18] dataset. The source prompt P_s for these videos is obtained using the caption model [13]. We develop the edit prompt P_e by adding or replacing some words.

5.2. Comparison with State-of-the-Art Methods

We compare our approach with four recent zero-shot methods, namely: FateZero [19], vid2vid-zero [30],

(a) Input **(b)** Ours **(c)** vid2vid-zero

Figure 5: Qualitative comparison of our method with vid2vid-zero. Best viewed with zoom-in

(a) Input **(b)** CG Style **(c)** Ghibli Cartoon **(d)** Ink Painting

Figure 6: More Results: Demonstration of editing for cartoon and painting styles

Pix2Video [4], and Text2Video-Zero [11], where the code for all the baselines is publicly available, but they generate plausible results only for editing templates mentioned in their code base. In particular, vid2vid-zero generated frames totally different from the input. For the sake of comparison, we have taken most of the videos that are common among these baselines. As shown in Figure 4, the case of editing "*swan*" to "*yellow terosaur*", this is mentioned by Fatezero as their limitation in their paper. As shown, we are able to successfully replace the "*swan*" to "*yellow terosaur*". Additionally, we compared this with Text2Video-Zero as well since this uses the ControlNet, hence, it has added information to edit the videos, but it generates frames that are not even close to the target prompt, instead, it changes the colour of the wall to yellow. Similarly, in Figure 5, we have compared our method against vid2vid-zero for the input video and target prompt mentioned in their zero shot results. The generated frames, as per their code base, show the backside of the "*Porsche car*" while moving forward, whereas in the input video, the Jeep is moving forward and is front-facing. We showed that our results are much better in terms of resemblance to input

(a) ***Source Prompt***: *Blooming field of red poppy flowers.* ***Target Prompt***: *Blooming field of white poppy flowers*

(b) ***Source Prompt***: *White Snowball Flowers.* ***Target Prompt***: *Cherry Blossom Flowers*

(c) ***Source Prompt***: ***Morning view*** *over a farm.* ***Target Prompt***: *Sunset view over a farm*

Figure 7: More Results: Demonstration of editing over fine-grained structure, shape and color

and target prompts. Following [19, 33, 5, 31] we also conducted the quantitative evaluation using the trained CLIP [20] model. Specifically, we show the "**Temporal**" [5] to measure the temporal consistency between consecutive frames by measuring cosine similarity between all pairs of consecutive frames. Another measure, "**Edit Acc**" [20, 31] to measure the editing accuracy in the generated frames, is calculated as the percentage of generated frames having a higher similarity for the target prompt than the source prompt. Additionally, we evaluated our method on two user studies metrics ('**Edit**' and '**Temporal**') are measured to measure the editing quality of our system from the perspective of the applicability of our method in terms of usage in a real environment. Specifically, we measured

Method	CLIP Metrics ↑		User Study ↓	
	Temporal	Edit Acc	Temporal	Edit
Tune-A-Video	0.934	0.738	2.79	2.73
vid2vid-zero	0.951	0.696	2.71	2.69
FateZero	0.954	0.894	1.89	2.52
Pix2Video	0.912	0.701	2.60	1.98
Text2Video-Zero	0.959	0.902	1.81	1.77
Ours	**0.971**	**0.915**	**1.78**	**1.31**

Table 1: Quantitative evaluation: For both user study and CLIP metrics INFUSION outperforms all the baselines in terms of temporal consistency and per frame editing accuracy

Input Ours w/o f_t^l-inject w/o att-inject w/o FUSION

Figure 8: Ablation Study of Feature Injection, Attention Injection and Attention Fusion. Prompts used are P_s: "*A **silver jeep** driving down a curvy road in the countryside*", P_e: "*A Porsche car driving down a curvy road in the countryside*". Without any of these components the foreground or background or both details are missing from the edited video.

the rank of our proposed method for temporal consistency ('**Temporal**') across the frames in edited video and overall frame-wise editing ('**Edit**') for a given target prompt. We asked 20 subjects to rank the editing method, with nine sets of comparisons in each study. As shown in Table 1, our proposed method INFUSION outperforms for all the CLIP metrics, hence achieving the best temporal consistence and better per-frame editing accuracy. Moreover, our method is truly zero shot since we have not used any other pre-trained diffusion other than Stable Diffusion v1.5[22] as opposed to **Fatezero**, which uses the one-shot trained model for the target prompt "*A Porsche car driving down a curvy road in the countryside*". Apart from CLIP metrics, our method is more reliable to put into real world editing since it earns user preferences the best among all methods in both (**Edit** and **Temporal**) aspects.

5.3. Ablation Study

Despite proving the effectiveness of INFUSION, in this section we will present the ablation study (shown in Figure 8) of various components in our editing method to discuss the importance of each and their contribution towards the edited video.

INJECTION is studied in Figure 8 as shown in the third and fourth columns. The column named "***w/o f_t^l-inject***" is the ablation study of feature injection, it depicts the complete change in frames including both foreground (*Porsche car*) and background (*curvy road and countryside as in source frames*). Though, we wanted to change the structure of "*jeep → Porsche car*", but without feature injection it changes the complete source layout/background as well, which is not a desirable change. The column named "***w/o att-inject***" is the ablation for attention injection. As shown, without attention injection, the rest of the background is faded, however, due to feature injection, the source layout/background are retained ("*curvy road and countryside*"). Additionally, due to feature injection, the target concept is highlighted ("*Porsche car*") but the remaining source layout is faded, and hence it is very much required to fill the source concepts on the highlighted target concepts, which is proposed to be done with self-attention injection since it can inject the source concepts while keeping the importance of highlighted target concepts injected using differential features.

ATTENTION FUSION is also studied in Figure 8, where we have removed self attention fusion (shown in the fifth column) as discussed in equation 9 but cross attention mixing is present. The resulting frames show little or no change, with some shape distortion in the front part of the Jeep, as if the diffusion is trying to align the Jeep structure with that of the car. As expected, the change injected using differential feature and attention until S_2 steps is wiped, and the concepts that are highlighted also get wiped due to continued diffusion steps without masking, and hence it will generate a structure similar to the source.

5.4. MORE RESULTS: FLEXIBLE STRUCTURE, COLOR AND STYLE

Our proposed method has shown decent results in editing the structure, colour, and shape at finer details (Figures 7 and 6) with source content preserved. Specifically, as shown in Figure 7a the source content contains lot of cluttered red flowers, which are edited to white flowers (fine-grained colour editing) with decent temporal consistency and accuracy over the frames. Similarly, Figure 7b contains a lot of snowball flowers, which are edited to "*cherry blossom*" flowers (fine-grained structure and shape editing) at all viewing angles demonstrated over frames. Figure 7c demonstrated fine-grained structure editing from "*morning*" view to "*sunset*" view with just deleting the sun and associated rays over the farm and remaining everything intact. Other results include style editing ranging from cartoon to painting styles, as shown in Figure 6. Here, as we see, the details like nails, pose, etc. of the woman are preserved. More results will be added to the github page: *https://infusion-zero-edit.github.io/*

Conclusion

In this work, we have presented the generalised zero-shot text-based video editing framework with no additional models like ControlNet[32] and with no training or fine-tuning of the pre-trained image diffusion model. Only pre-trained Stable Diffusion v1.5[22] is used for all the edited videos without the need for any customised image diffusion.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022.

[2] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023.

[4] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. *arXiv preprint arXiv:2303.12688*, 2023.

[5] 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.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] 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.

[8] 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.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[10] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022.

[11] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.

[12] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[14] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023.

[15] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[16] 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.

[17] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[18] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[19] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

[20] 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. PMLR, 2021.

[21] 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.

[22] 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.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[24] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[25] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*, 2023.

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

[27] Raphael Tang, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022.

[28] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 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*, pages 1921–1930, 2023.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023.

[31] 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.

[32] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[33] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023.