Investigating the Effectiveness of Cross-Attention to Unlock Zero-Shot Editing of Text-to-Video Diffusion Models

Saman Motamed[1] Wouter Van Gansbeke[1] Luc Van Gool[1,2,3]

[1] INSAIT, Sofia University, Bulgaria [2] ETH Zurich [3] KU Leuven

Abstract

With recent advances in image and video diffusion models for content creation, a plethora of techniques have been proposed for customizing their generated content. In particular, manipulating the cross-attention layers of Text-to-Image (T2I) diffusion models has shown great promise in controlling the shape and location of objects in the scene. Transferring image-editing techniques to the video domain, however, is extremely challenging as object motion and temporal consistency are difficult to capture accurately. In this work, we take a first look at the role of cross-attention in Text-to-Video (T2V) diffusion models for zero-shot video editing. While one-shot models have shown potential in controlling motion and camera movement, we demonstrate zero-shot control over object shape, position and movement in T2V models. We show that despite the limitations of current T2V models, cross-attention guidance can be a promising approach for editing videos. Code: https://github.com/sam-motamed/Video-Editing-X-Attention.git

1. Introduction

Text-to-Video diffusion models [2, 3, 18, 41, 46] have been fast advancing in generating temporally consistent scenes with plausible object interactions. There has been a series of works that have focused on editing T2V models to enable greater control over video generation. More successful editing methods have made use of small sets of reference videos to learn an object's motion or camera movement. They subsequently transfer that specific movement or camera motion to a new object and scene [21, 52, 56] by training parts of the video diffusion model or performing Low Rank Adaptation (LoRA) [19]. While these methods can be effective, they require additional data and compute with limited flexibility, which limits their adoption in practice.

Several works [40, 45, 47, 49] have shown the promise of attention maps in object discovery and segmentation. In the domain of text-to-image models, cross-attention and its role in controlling the scene layout has also been well studied. In particular, cross-attention is responsible for determining the objects' shape and size in the image. Cross-attention facilitates maintaining semantic consistency between the text and the generated image. By attending to relevant textual features, the model ensures that the generated visual content aligns with the overall semantics of the input description. One of the works that exploited cross-attentions to enable editing images was Prompt-to-Prompt [16]. This work showed that the shape of an object **a** can be replaced with the shape of another object **b** by replacing **a**'s cross-attentions with those of **b**. Training-Free Layout Control [7] was another work that proposed an energy-based objective to control the position of objects in the generated image. Given a user-specified bounding box, the energy function encourages the cross-attention maps of a token to form within the bounding box and hence position the object within the bounding box. Diffusion self-guidance [12] generalized the Training-Free Layout Control method such that editing the scene could be done by using the cross-attention maps alone, without the need for external inputs (e.g., bounding box), in a zero-shot manner. This was achieved by applying transformations (e.g., relocating and resizing) to the original cross-attentions of a token and using the resulting cross-attentions as the target of the objective.

With the success of the above methods for editing images generated by T2I models [5, 12, 16, 30] or adapting T2I models for video editing [27], we ask the question; "Do such approaches for editing images transfer to the video domain?". In particular, we are interested in exploring the effectiveness of cross-attention layers for editing the subject's size, positioning, and motion in videos.

In this paper, we build upon the achievements of prior image-editing techniques by extending them to the video domain. More specifically, our contributions are threefold:

- We take a first look at cross-attention layers in T2V diffusion models and their role in editing videos.
- We explore two possible ways to use cross-attentions in editing videos; namely *forward* and *backward* guidance.

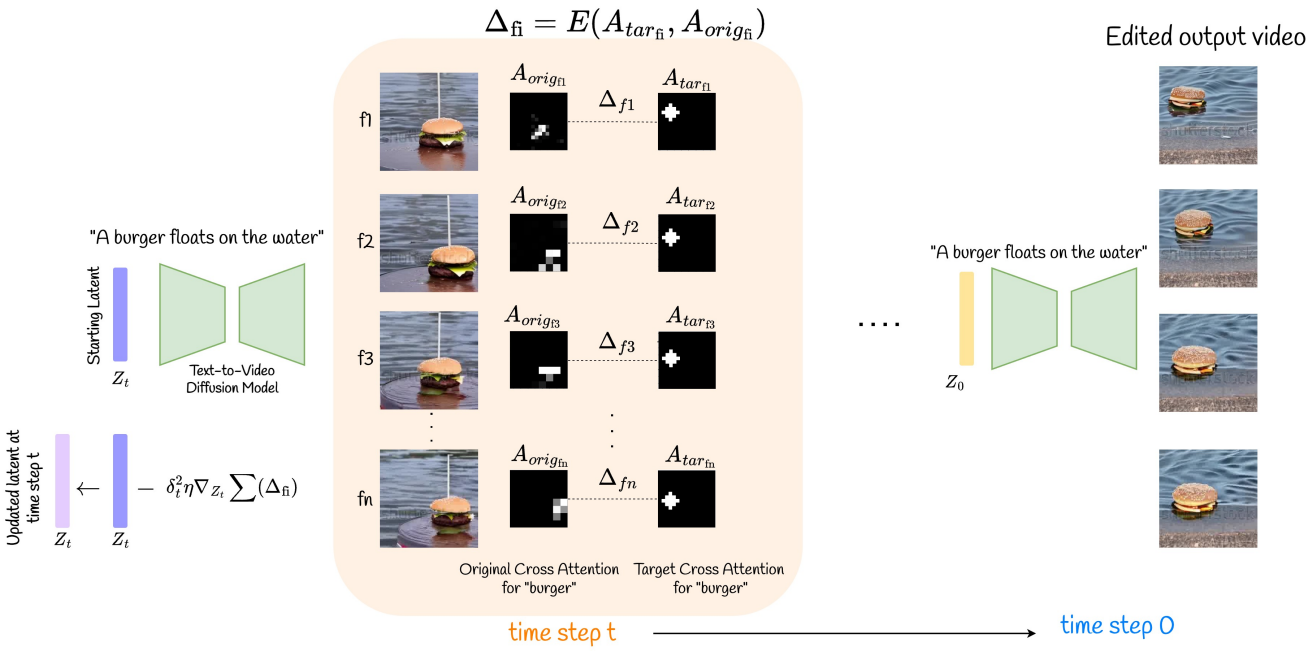

Figure 1. This figure shows an overview of backward guidance in T2V models. On the left, we show the generated frames of the T2V model after t steps, given an initial input latent z_t and the text prompt "A burger floats on the water". To edit the video and move the burger from the top-left of the screen to the bottom-left in a straight line, we generate $\mathcal{A}_{tar_{\text{fi}}}$ for each frame **fi** reflecting this edit. Following the scheme in Section 4.2, we update the latent through the denoising process based on objective E. At time step 0, z_0 generates the video on the right which reflects the intended edit.

- We investigate the limitations of current T2V models that hinder the capabilities of video editing methods.

2. Related Works

Denoising Diffusion Models. The denoising diffusion paradigm [1, 42, 43] emerged as a new method to generate images with high photo-realism and diversity. It has rapidly advanced text-conditioned image generation [11, 14, 17, 31, 34–36, 39, 54, 55], which is important for gaining control over its generated content. Due to their versatility and representation learning capabilities [8, 20, 53], they have also been successfully adapted for specialized tasks such as classification [10, 26], depth prediction [23] and segmentation [6, 22, 44].

Personalizing Image Generation. Personalizing [13, 25, 29, 38, 50, 55] and editing [5, 12, 16, 30, 32] T2I models has become a research focus to enable user-intuitive control for creating content with these generative models. In particular, the cross-attention layers of diffusion models have been studied for their role in determining a scene's layout and their ability to enable zero-shot editing of generated images. Similar to [5], we split cross-attention-based editing of T2I and T2V models into two categories of 1) **forward** and 2) **backward** guidance.

In **forward** guidance, cross-attention manipulation occurs directly during the denoising process via a forward pass through the model. A notable example of forward guidance is Prompt-to-Prompt [16], which proposes replacing the token's cross-attentions from a source prompt with those of a target prompt. Figure 2 shows one such example in the video domain where the cross-attentions of "car", from the source prompt "car drives on the road", are replaced with cross-attentions of "truck", from the target prompt "truck drives on the road". To enable more precise modifications to a specific source token, while preserving the overall scene, forward guidance requires source and target prompts that differ by a single token, limiting its applicability.

In contrast to forward guidance that directly manipulates cross-attentions, **backward** guidance biases the cross-attention through backpropagation. By designing an energy-based loss that encourages some desired edit [5, 12], the gradient of the loss is then used to update the input latent z_t of the model. Training-Free Layout Control [5] is an example of backward guidance where the energy function encourages the cross-attentions of the user-specified token to obtain higher values inside a user-defined bounding box. At multiple time steps, the input latent is updated to realize this objective. Similarly, Diffusion self-guidance [12] de-

signed energy functions that encourage the cross-attentions to take certain shapes or positions within the image. This paper is inspired by the success of these two works in the image domain. In Section 3.3, we show that forward guidance is too restrictive to enable effective video editing. In Section 3.4, we show backward guidance's promise in enabling zero-shot editing of T2V models.

Text-to-Video Generation. Diffusion models have been improving at high-quality video generation by training conditional denoising networks (e.g, 3D U-Net [9], DiT [33]) to denoise randomly sampled sequences of Gaussian noises [2, 3, 18, 28, 41, 46]. Some works take advantage of large, pre-trained text-to-image foundation models to build text-to-video models. This is done by inflating the T2I model with temporal layers, like Tune-A-Video [51], Text2Video-Zero [24] and AnimateDiff [15].

Personalizing Video Generation. Following the same desire to control image generation, a few works focused on video editing and customizing the motion and camera movement in T2V models [7, 21, 48, 52, 56]. Most current editing and customization methods work by tuning parts of the network or performing LoRA [19] based on example videos containing the desired effect. Such methods lack the flexibility of a zero-shot approach and require additional training data and resources. For this reason, we investigate the effectiveness of forward and backward guidance using cross-attention for T2V models.

3. Method

3.1. How Do Video Diffusion Models Work?

Video diffusion models train a 3D denoising network, traditionally U-Nets but more recently transformer-based [33] networks, to generate videos from randomly sampled Gaussian noise. In this work, we use T2V models with 3D U-Net backbone [46] which consists of down-blocks, middle-blocks, and up-blocks. Each block has several convolution layers, spatial transformers, and temporal transformers. During training on videos, the U-Net (ϵ_θ) and a text encoder (τ_θ) are optimized with the following objective:

$$\mathcal{L} = \mathbb{E}_{z_0, y, \epsilon \sim \mathcal{N}(0,I), t \sim \mathcal{U}(0,T)} = \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(y))\|_2^2, \tag{1}$$

where $z_0 \in \mathbb{R}^{f \times b \times h \times w \times c}$ is the initial latent input of the training videos (b indicates the batch size, f is the number of frames, h, w and c are the height and width and channels respectively) and y is the text description of the video, with ϵ and t being the added Gaussian noise to the videos and the time step. At time step t, the noised latent is defined as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{2}$$

where α_t controls the noise strength.

3.2. T2V Cross-attention

The cross-attention mechanism in the spatial transformers of the 3D U-Net enables the model to capture spatial relationships between the video frames and the input text. In this work, we focus on changing an object's size, location and motion given a latent input and text prompt to the T2V model. To this end, we work with the cross-attention layers of the 3D U-Net where $\{\mathcal{A}_{i,t,.,.,k} \in \mathbb{R}^{H_i \times W_i \times |k|}\}$ is the Softmax-normalized cross-attention map of the i^{th} layer of the U-Net, at time step t for token k.

3.3. Forward T2V Guidance

Following the works that perform forward guidance in T2I models [4, 16, 32], we implemented forward guidance in the T2V pipeline. Figure 2 is one example where the cross-attentions of "car" are replaced with the cross-attentions of "truck". Below are the two main limitations with forward guidance that have also been observed in the T2I domain.

- **Size and Shape Mismatch**. Forward guidance is restrictive and can lead to artifacts due to the difference in shape and size of the two objects. In the example of Figure 2, since the truck is larger than the car, injecting the cross-attentions of the truck to replace the car's has led to artifacts around the car without changing the car's size to match the truck's.
- **Cross-attention Overlap**. The cross-attentions of different tokens can overlap. We refer to the top row of Figure 3, where the shark is still visible in the cross-attention maps of tokens "in" and "the". For this reason, forward guidance can work reasonably well where the two source and target sentences only differ by one token (i.e., Prompt-to-Prompt's setting). This overlap can cause degradation in the image and video quality, especially when the text inputs differ by more than one token.

We note that some of these artifacts are due to the current T2V models generating noisy cross-attentions. We go over more details in Section 4.1 regarding this limitation.

3.4. Backward T2V Guidance

Following Diffusion self-guidance [12] and Training-Free Layout Control [5], we define an energy function E to encourage specific shape, size and motion properties on the cross-attentions of some user-specified token k. Figure 1 gives an overview of our backward guidance where $\mathcal{A}_{orig_{\text{fi}}}$ is the cross-attention map of some user-specified token k (e.g., token corresponding to "burger") in frame fi of the video generated by the T2V model. we omit the layer number and the token k in our notation of the cross-attention. $\mathcal{A}_{tar_{\text{fi}}}$ is the target cross-attention that captures the properties of the editing task. In Figure 1, the task is to move the burger from the top-left to the bottom-left of the scene. We

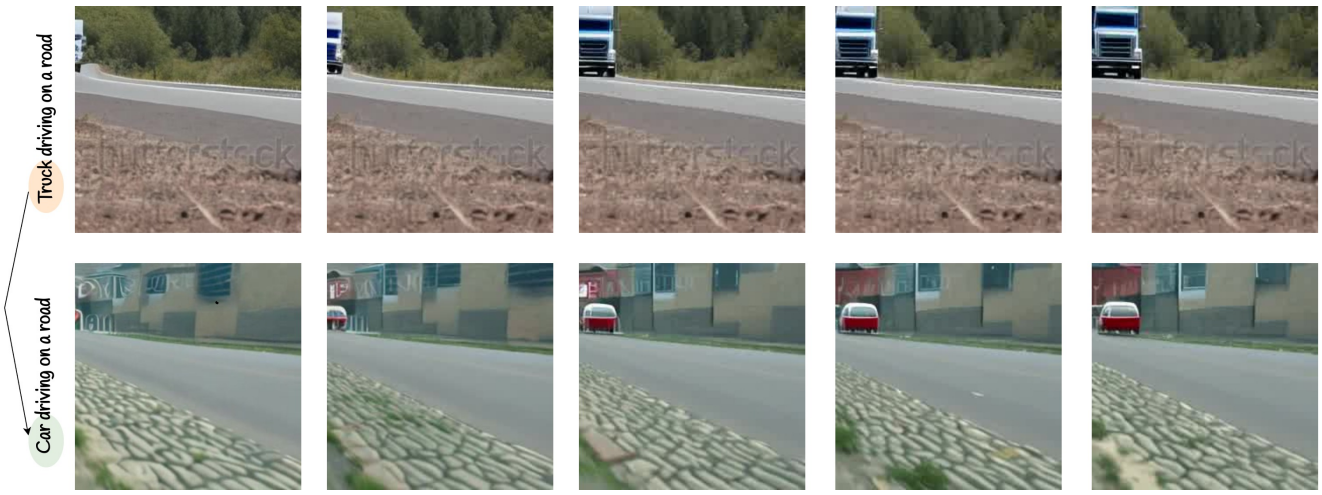

Figure 2. We show an example of forward guidance by swapping the cross-attention maps of "car" with cross-attention maps of the "truck". The two input texts only differ in one token ("truck" and "car"). While the car follows the motion and location of the truck in the video, artifacts can be seen around the car due to the mismatch in size and shape of the truck and car.

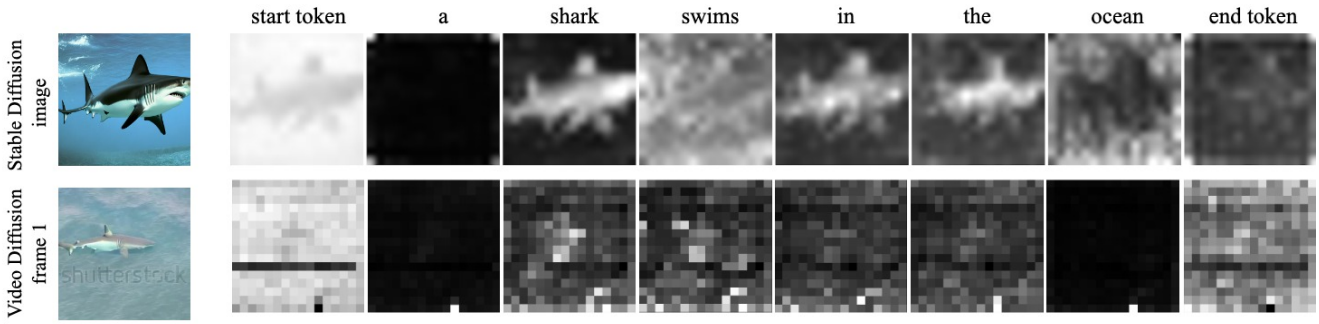

Figure 3. We compare the cross-attention maps for the same prompt to a T2I and T2V model. The cross-attention maps are extracted and averaged at the 16×16 resolution from the mid-blocks and up-blocks of the U-Net. Open-source T2I models currently produce much less noisy cross-attention maps compared to T2V models. In Section 4.1, we give details on how the noisy cross-attentions hinder backward guidance and propose a procedure for bypassing this limitation for our experiments in this paper.

define the energy function E below. To control the shape and size of an object (indicated by token k) through its corresponding cross-attention maps, we threshold the attention map to eliminate the effect of background noise and overlapping attention from other tokens. This is achieved by taking a soft threshold at the midpoint of the per-channel minimum and maximum values:

$$\text{shape}(k) = \mathcal{A}_k^{\text{threshold}}.$$

Using the thresholded original cross-attention and the target cross-attention, we define the energy function E as:

$$E = \text{shape}(A_{tar}) - \text{shape}(A_{orig}). \tag{3}$$

This objective is zero-shot since $\text{shape}(A_{tar})$ can be computed as $(M \times \text{shape}(A_{orig}))$ where M defines some transformation such as resizing and relocating the original attention. At time step t, we update the latent z_t according to the gradient of the loss defined by the energy function E. This is realized through the following equation:

$$z_t \leftarrow z_t - \delta_t^2 \eta \nabla_{Z_t} \sum E(A_{tar}, A_{orig}), \tag{4}$$

where $\eta > 0$ controls the strength of backward guidance and $\delta_t = \sqrt{(1-\alpha_t)/\alpha_t}$. Updating the latent z in this manner indirectly influences the cross-attentions. Please refer to Section 4.2 for more details on our experimental setup.

4. Experiments

4.1. Limitation of Current T2V Models

In Figure 3, we visualize the cross-attention maps for all tokens of the prompt "a shark swims in the ocean" generated with Stable Diffusion [37] and our T2V model [46].

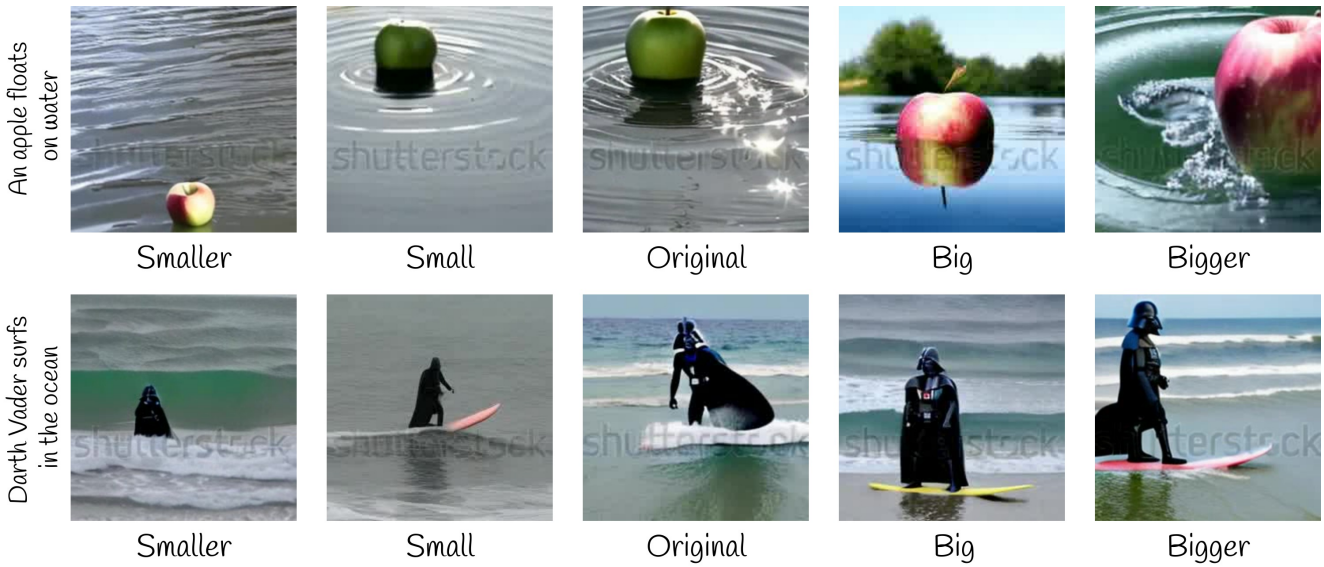

Figure 4. We show qualitative results for shrinking and enlarging objects through backward guidance. The middle image of each row visualizes the first frame of the original video. We enlarge and shrink the target cross-attentions at four different levels (Big / Bigger and small / smaller) and update the latent through backward guidance. The first frame for each edited video is shown.

The cross-attention maps in T2I models capture the tokens much better than T2V models. We attribute this to deeper denoising networks of T2I models, larger training datasets, and more cross-attention layers. Using such noisy cross-attention maps hinders both forward and backward guidance. To perform backward guidance more effectively, we opted to directly generate shape(A_{tar}) for each frame. Instead of transforming shape(A_{orig}) to calculate shape(A_{tar}) for each video frame, we generate binary cross-attention maps for the token of interest. Despite this backward guidance setup not being zero-shot, we rely on future T2V models with better cross-attention maps to replace this manual effort.

Figure 1 shows an example of user-generated target cross-attentions. In this example, instead of transforming the cross-attention maps of the "burger" to calculate the target, we directly generate each frame's cross-attention according to our editing task. Here, the task is to move the burger from the top-left of the scene to the bottom-left in a straight line. Hence, we generate cross-attention maps for each frame. For frame 1, $A_{tar_{\text{fl}}}$ is placed at the top-left of the scene and in the following frames, the cross-attention map moves slightly down such that in the last frame, $A_{tar_{\text{fl6}}}$ is placed at the bottom-left.

4.2. Experiment Details

We use the ModelScope [46] T2V model in our experiments and generate 16 frame videos with 256×256 resolution. Image editing methods such as Diffusion self-guidance [12] have used the extracted image features learned by the denoising network to preserve the background details and appearance features of the object being edited. In this work, we only focus on controlling objects' motion and size and leave background and appearance consistency for future works. We experimented with text prompts that describe a simple scene, to further control the limitation of current T2V models and get less noisy cross-attention maps for the object we want to edit.

Our 3D U-Net has cross-attentions with resolutions 4×4, 8×8 , 16×16 and 32×32. We find the 8×8 and 16×16 cross-attentions to be the most important dimensions for effectively minimizing the energy function and editing the scene. There are 10 such layers in the down-blocks, mid-blocks, and up-blocks of the 3D U-Net (4 down-block, 2 mid-block, and 4 up-block layers). We found that mid-block's cross-attentions played a vital role in backward guidance. Excluding the two mid-block layers resulted in failed edits whereas excluding either all of down-block's or all of up-block's cross-attentions resulted in fewer failures.

We experimented with different schemes for updating the latent z and found the most effective strategy to be that of Diffusion self-guidance [12]. During the first $N/4$ iterations, we update z at each step. For the subsequent $3N/4$ iterations, we update z at every other step. The guidance scale η (eq. 4) also plays an important role in the method's effectiveness. Increasing η too much leads to degradation in the generated frames. Selecting a very low scale does not change the latent enough for effective editing. We found that in our setting, a scale of $15 < \eta < 25$ provided a good balance between guidance strength and synthesis quality.

5. Results

In this section, we show the capabilities of backward guidance for two different tasks. Figure 4 shows qualitative results for changing an object's size through backward guidance. Figure 5 presents qualitative results of backward guidance for controlling the motion of an object in a video. To edit an object of interest, we generate binary cross-attention maps that capture the target position for the object's token. For "burger", we placed the first cross-attention at the top-left of the scene and slowly moved it down. For the "ball", we placed the cross-attention at the top-left and moved it towards the bottom-right of the scene. Finally, we moved the "shark" from the top-right towards the bottom-left of the scene. Each sequence of frames with the black caption shows the original video without performing guidance. The sequence of frames with blue instruction shows the video after updating the latent with backward guidance. The object successfully follows the cross-attention at each frame.

We also observe that the original video can be missing an object described in the text. The example with prompt "A wolf howls to the moon" in Figure 5 is missing the moon. Interestingly, backward guidance encourages the moon to be present in the scene. Attend-and-Excite [4] achieves the same objective in the T2I domain.

6. Observations

In this section, we go over a few interesting observations when experimenting with backward guidance.

Perspective. In our experiments, we used a fixed size for A_{tar} for all 16 video frames. However, if the object is moving away or toward the camera, we should see a change in the object's size. In Figure 5, we see the burger, ball, and shark getting larger as they move closer to the camera while the moon remains the same size as it is static in the sky. It is noteworthy that despite updating the model's input latent with fixed-size target cross-attentions, the model consistently generates videos with reasonable perspective. However, this comes at the expense of not strictly adhering to the exact size defined by A_{tar}.

Motion Control. To control the motion of objects, we interpolate the cross-attention maps between the attention map $A_{tar_{\text{fl}}}$, placed at starting position a and the attention map of the last frame $A_{tar_{\text{fl6}}}$ placed in final position b. We observed that the model keeps the temporal consistency at the expense of not following the exact start and end location defined by the target cross-attention. For instance, in Figure 5 - last row, we placed the cross-attention of the "shark" at the top-right for the first frame and at the bottom-left for the last (16^{th}) frame. However, after t steps, the shark is not at the bottom of the scene where $A_{tar_{\text{fl6}}}$ was positioned. To do so, the model needs to move the shark much faster to go from $A_{tar_{\text{fl}}}$ to $A_{tar_{\text{fl6}}}$ in a short number of frames. We also note that compared to resizing an object, controlling its motion is often prone to failures using backward guidance. This failure takes the form of the object being statically positioned at $A_{tar_{\text{fl}}}$. We leave further exploration of this mode of failure for future work.

7. Discussion

This study conducted an initial investigation into the significance of cross-attention layers within the 3D U-Net framework of video diffusion models. More specifically, focusing on their role in determining objects' size, position and motion in T2V models. We examined the efficacy of utilizing cross-attention maps to manipulate object size and motion, employing both forward and backward guidance. In Section 3.3, we showed that forward guidance in videos faces the same limitations that were previously observed in the T2I domain [7] which hinders its performance. In Section 5, we showed results for editing the size and motion of an object through backward guidance. Our findings emphasize the promise of backward guidance in enabling zero-shot editing capabilities for video generation. Moreover, in Section 4.1, we highlighted current limitations that impede the transition of cross-attention-based editing methods from the image domain to videos. This analysis provides insights into the challenges and opportunities inherent to adapting editing techniques to be used in dynamic video content.

8. Impact and Future Directions

Enabling zero-shot editing capabilities for generative video models is a valuable approach to enhance user control without the need for model fine-tuning with additional data. While current video models face limitations in quality, length, and cross-attention accuracy, we anticipate that editing methodologies like ours will leverage future advancements in Text-to-Video models, similar to the progress seen in the Text-to-Image domain.

In this study, we focused on manipulating objects' size and motion with backward guidance. However, practical applications for editing tools require further exploration, particularly enabling editing of real videos. This needs additional constraints such as controlling background alterations and maintaining the fidelity of different objects to the original video. These aspects remain open for future work.

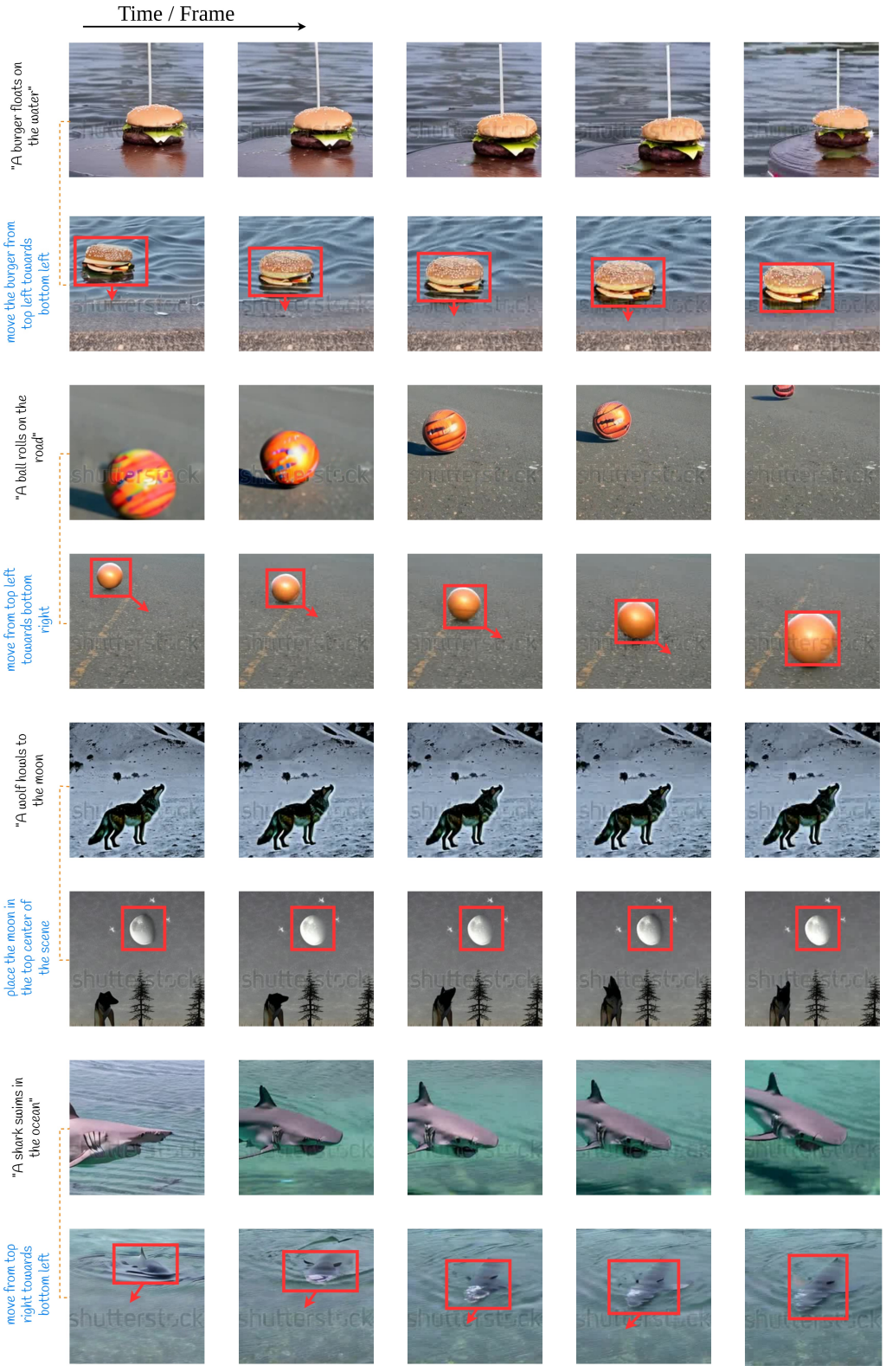

Figure 5. The figure visualizes the results of backward cross-attention guidance. For each of the 4 examples, we show the output of the T2V model given the prompt in black. The blue text describes the applied transformation to the cross-attentions at each frame. We update the input latent accordingly. The red bounding box highlights the edit's success.

References

[1] Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 1, 3

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1, 3

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3, 6

[5] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024. 1, 2, 3

[6] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 909–919, 2023. 2

[7] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 1, 3, 6

[8] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning, 2024. 2

[9] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 3

[10] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[12] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 5

[13] 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. 2

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[15] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[16] 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. 1, 2, 3

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[18] 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

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 3

[20] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779*, 2023. 2

[21] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. *arXiv preprint arXiv:2312.00845*, 2023. 1, 3

[22] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation, 2023. 2

[23] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[24] 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. 3

[25] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2

[26] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023. 2

[27] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control, 2023. 1

[28] Kangfu Mei and Vishal Patel. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9117–9125, 2023. 3

[29] Saman Motamed, Danda Pani Paudel, and Luc Van Gool. Lego: Learning to disentangle and invert concepts beyond object appearance in text-to-image diffusion models. *arXiv preprint arXiv:2311.13833*, 2023. 2

[30] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 1, 2

[31] 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. 2

[32] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models, 2023. 2, 3

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[34] 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. PMLR, 2021. 2

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

[36] 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. 2

[37] 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. 4

[38] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[40] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021. 1

[41] 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

[42] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[44] Wouter Van Gansbeke and Bert De Brabandere. A simple latent diffusion approach for panoptic segmentation and mask inpainting. *arXiv preprint arXiv:2401.10227*, 2024. 2

[45] Wouter Van Gansbeke, Simon Vandenhende, and Luc Van Gool. Discovering object masks with transformers for unsupervised semantic segmentation. *arXiv preprint arXiv:2206.06363*, 2022. 1

[46] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 3, 4, 5

[47] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *CVPR*, 2023. 1

[48] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[49] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *CVPR*, 2022. 1

[50] Chen Henry Wu, Saman Motamed, Shaunak Srivastava, and Fernando D De la Torre. Generative visual prompt: Unifying distributional control of pre-trained generative models. *Advances in Neural Information Processing Systems*, 35:22422–22437, 2022. 2

[51] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3

[52] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion, 2024. 1, 3

[53] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18938–18949, 2023. 2

[54] 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 autoregres-

sive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023. 2

[56] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 1, 3