Making Video Models Adhere to User Intent with Minor Adjustments

Anonymous authors
Paper under double-blind review

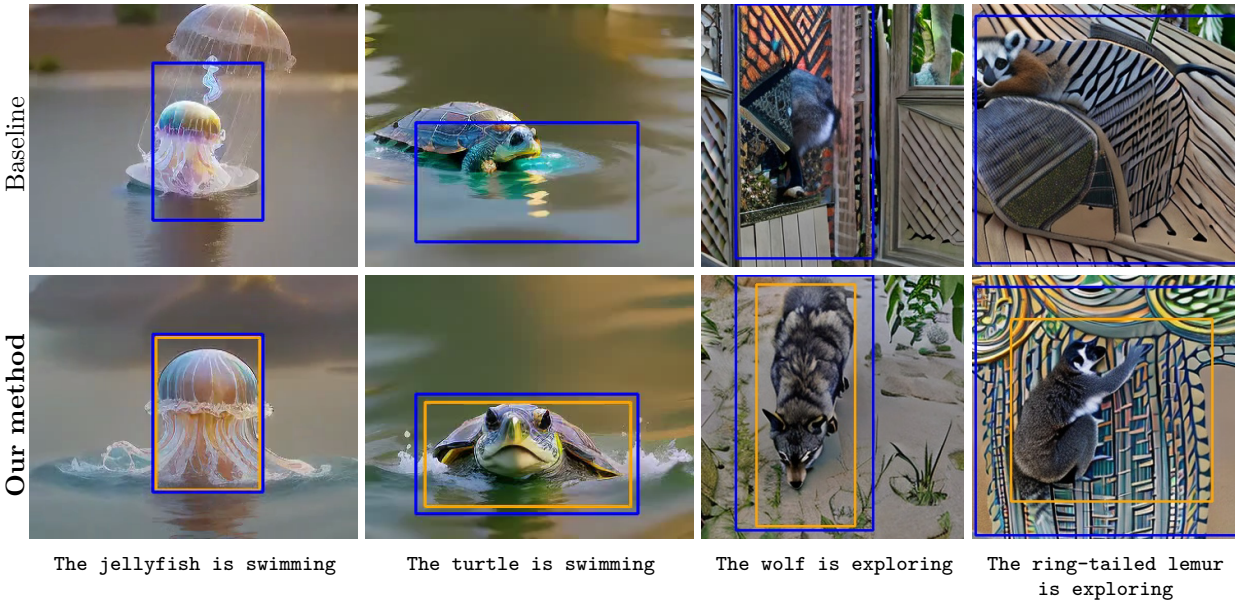


Figure 1: **Teaser** – We show example bounding box controlled video generations with (left) Trailblazer Ma et al. (2024b) in T2V-Turbo Li et al. (2025) and (right) the original Trailblazer Ma et al. (2024b) respectively. On top is the original control signal and below is our adjusted bounding boxes. We show the original bounding boxes as blue and the adjusted boxes as orange. While the modification is subtle, the difference in the quality of generation is large. We modify bounding boxes to adhere better to the cross-attention maps within the video models.

Abstract

With the recent drastic advancements in text-to-video diffusion models, controlling their generations has drawn interest. A popular way for control is through bounding boxes or layouts. However, enforcing adherence to these control inputs is still an open problem. In this work, we show that by slightly adjusting user-provided bounding boxes we can improve both the quality of generations and the adherence to the control inputs. This is achieved by simply optimizing the bounding boxes to better align with the internal attention maps of the video diffusion model while carefully balancing the focus on foreground and background. In a sense, we are modifying the bounding boxes to be at places where the model is familiar with. Surprisingly, we find that even with small modifications, the quality of generations can vary significantly. To do so, we propose a smooth mask to make the bounding box position differentiable and an attention-maximization objective that we use to alter the bounding boxes. We conduct thorough experiments, including a user study to validate the effectiveness of our method.

1 Introduction

Text-to-video diffusion models have made groundbreaking advances in producing high-quality prompt-directed generations Ho et al. (2022b); Singer et al. (2023); Weissenborn et al. (2020); Arnab et al. (2021); Ho et al. (2022a); Chen et al. (2024a). Among the various directions, methods based on diffusion Ho et al. (2022b;a); Chen et al. (2024a) and transformers Weissenborn et al. (2020); Arnab et al. (2021) have become popular. While these models can be controlled through proper prompting, this is not always straightforward and require careful prompt engineering Liu et al. (2023). In particular, the spatial control of object placement and object trajectories remains difficult.

Naturally, researchers have sought to improve the controllability of text-to-video diffusion models. These include methods that specifically train a model in addition to the main model Zhang et al. (2023), which was shown to be especially effective for text-to-image generation Wang et al. (2023); Zhang et al. (2023); Patashnik et al. (2021). For video generation, however, training such additional models is computationally expensive. For spatial control, using bounding boxes or layouts as control inputs Ma et al. (2024b); Zheng et al. (2023) has gained attention. These approaches work without additional training by simply modifying the internal attention maps within the video diffusion model Ma et al. (2024b); Hertz et al. (2023); Chefer et al. (2023); Tumanyan et al. (2023) or through guidance Patashnik et al. (2021); Nichol et al. (2022) with a pre-trained classifier. While these methods are effective, they are still limited in terms of adherence to the control inputs. Generated videos contain artificial outcomes because of the mismatch between how the control signals affect the video generation process and how it was trained without such injection; see Figure 1.

In this work, we show that by slightly adjusting the control inputs, we can improve both the quality of generations and the control adherence. However, finding how to adjust bounding boxes to adhere to user control without hurting the final generation outcome is challenging. For instance, modifying and balancing attention within the diffusion model is effective for control but can lead to over-saturation. In general, changes that are outside of the model's internal understanding easily cause video generation to degrade. To address these, we optimize bounding boxes to better align with the internal attention maps of the video diffusion model, while still being close to the user-provided bounding boxes. More specifically, we introduce an optimization framework that ensures that attention maps edits remain differentiable with respect to the bounding box parameters. With the differentiable pipeline, for improved alignment of the bounding box control and the video model, we then propose to maximize the attention within the bounding box of the *next layer after the edit*, which is representative of where the neural network is focusing on once the edits are applied. Further, while we enhance focus, we also balance attention between the foreground and background areas, so that generation process does not completely ignore the background. While the resulting adjustments to the bounding boxes are small, their impact on the video generation is significant; see Figure 1.

To summarize, we make the following contributions:

- we demonstrate that small adjustments to user intent can lead to significant improvements in controlled video generation when adjusting with our method;
- we propose a novel method to optimize the control inputs to align well with the internal attention maps of the video diffusion model, while still being close to the user-provided bounding boxes;
- to do so, we present an editing pipeline that alters a non-differentiable method to be differentiable with respect to the bounding box parameters;
- to find better bounding-box parameters, we propose a balanced attention maximization objective considering the cross-attention maps of the next layer after the edits; and
- we conduct experiments including a user study to validate the effectiveness of our method, and show that it outperforms existing methods in terms of video generation quality and adherence to user intent.

2 Related Work

Text-to-image generative models. Generative text-to-image models have shown ground-breaking results in terms of high-quality images Rombach et al. (2022); Nichol et al. (2022) synthesizing different objects,

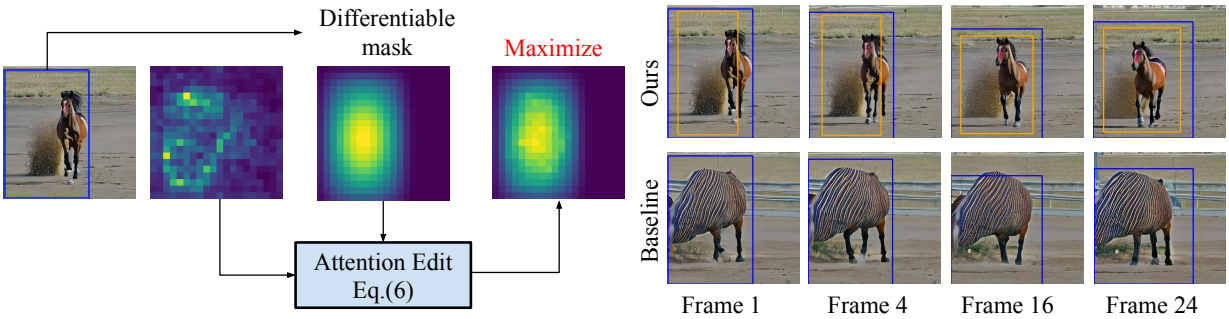


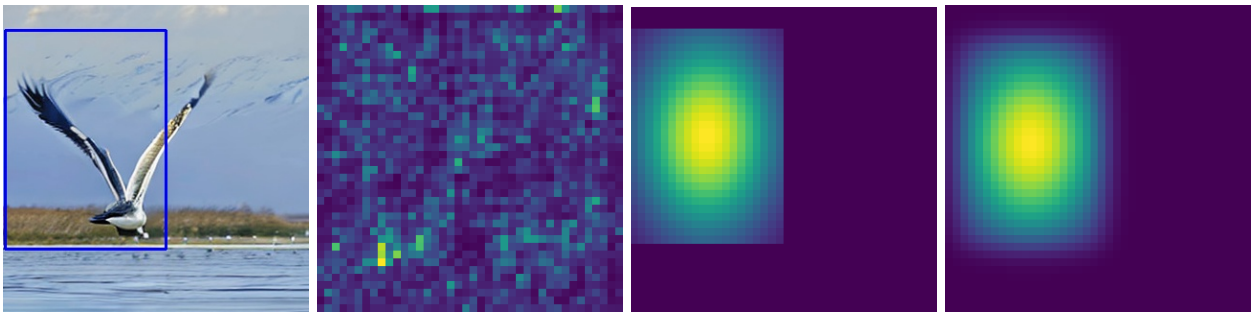
Figure 2: **Overview** – We inject bounding box control for video diffusion models by editing their cross attention maps within the network. However, not all such edits are friendly to video diffusion models as they are not trained with such edits. Thus, when applying these edits, we make sure that this editing process is differentiable (Section 3.2) and adjust the edit parameters in a way such that the network behaves as intended—attention being focused on desired regions (Section 3.3). We show the original bounding boxes in blue and the adjusted bounding boxes in orange. Though the adjustments are minimal and close to the original user input, they create a drastic difference in terms of video generation quality and adherence to the bounding boxes.

persons, and places. However, these models have been shown to fail to adhere faithfully to spatial user intent Rombach et al. (2022); Ramesh et al. (2022); Nichol et al. (2022), hence requiring a separate modality e.g., boxes, scribble, etc., for better text-to-image alignment.

Text-to-image generative models with box control. With alternate input control such as simple user-defined boxes, several methods Zheng et al. (2023); Chang et al. (2023) have demonstrated that the spatial composition can be controlled more faithfully. For example, Directed diffusion Ma et al. (2024a) directs the placements of objects by introducing activations at desired positions in a text-to-image diffusion model, leading to a better generation outcome. BoxDiff Xie et al. (2023) introduces spatial constraints such as inner-box, outer-box, and corner constraints in an optimization framework, for controlling objects and context in generated images. These methods follow a forward guidance approach. Alternatively, other work Chen et al. (2024b) demonstrates the superiority of backward guidance over forward guidance for robust layout control. Their method optimizes the latent, allowing both guided and unguided tokens to influence the generation outcome. However, these methods are limited to single images and do not investigate their effectiveness on temporal data, i.e., video.

Text-to-video generative models with box control. There are several works that investigate the adherence to input control for videos Wang et al. (2023); Lian et al. (2024); Jain et al. (2024); Ma et al. (2024b); Wang et al. (2024); Chen et al. (2025); Luo et al. (2025); Qiu et al. (2024); Lei et al. (2025). None, however, look into how slight changes in controls can lead to diverse final generation outcomes. Peekaboo uses attention masking to guide a video generation process, therefore utilizing local context for generating individual objects. However, their use of an infinite attention injection in the background often results in missing background details Ma et al. (2024b). Trailblazer, on the other hand, uses a direct/in-place replacement strategy to bias cross-attention maps towards user intent. With keyframed boxes, their method achieves controllable generation but often sacrifices better generation outcomes for more control. In other words, controlled generation outcomes are less faithful to the given prompt. Their method also relies on tuning hyper-parameters per object, which is not scalable.

To the best of our knowledge, we are the first to explore the benefit of adjustments to box-controlled generation outcomes. We build on the work of Trailblazer Ma et al. (2024b), and apply our method to two text-to-video models Ma et al. (2024b); Li et al. (2025). We demonstrate our core contributions in the following sections.



(a) Generated frame and user control (b) Attention map for the token 'Swan' (c) Attention map edited by Trailblazer Ma et al. (2024b) (d) Attention map edited by our method

Figure 3: **Example of attention map editing** – We show an example of the generated video frame with the desired user control bounding box, and the associated attention map edits. While effective, Trailblazer Ma et al. (2024b) relies on a replacement operation that is not differentiable with respect to the box parameters. Our method, on the other hand, performs a smooth differentiable edit.

3 Method

Our core idea is to modify user input, i.e., the bounding boxes, aligning them to the internal attention mechanisms within the video diffusion model. As shown in Figure 2, we implement this by introducing a differentiable attention map editing method, which we then optimize through to adjust the bounding boxes. When adjusting, as we do not have any specific measure for what a good trajectory is—after all, we are generating a video from scratch—we look into how the attention maps evolve through the network layers. Specifically, we encourage the edits to be inline with what how the neural network creates further attention maps, that is, we regularize it such that the attention map, after editing, causes the attention map of the next layer within the neural network to focus within the box.

To explain our method, and for completeness, we first review how control injection is done for a training-free baseline, then discuss our work.

3.1 Preliminary: Trailblazer Ma et al. (2024b)

Our goal is to generate a video with F frames, where a desired object of interest, e.g., "cat", adheres faithfully to the motion and control of the bounding boxes $\mathcal{B} \in \mathbb{R}^{F\times 4}$ provided by the user. We build upon Trailblazer Ma et al. (2024b), which 'injects' user control in the form of bounding boxes, by directly adjusting the internal attention maps of the video diffusion model. A strong benefit in doing so is that there is no need to train or fine-tune the video diffusion model, and any off-the-shelf video diffusion model can be used as long as it has cross-attention layers.

Specifically, at inference, we feed an input text embedding $\boldsymbol{p}$, noisy latent code $\boldsymbol{z}$, and timestep t, as input to a video diffusion model Θ_b. After each cross-attention layer, we extract and edit cross-attention maps $\boldsymbol{A}_S \in \mathbb{R}^{C\times H\times W\times N}$, temporal cross-attention maps $\boldsymbol{A}_T \in \mathbb{R}^{C\times H\times W\times F\times F}$. Then, denoting the predicted noise sample as ϵ_t, we can write

$$\boldsymbol{A}_S, \boldsymbol{A}_T, \epsilon_t = \Theta_b(\boldsymbol{p}, \boldsymbol{z}, t), \tag{1}$$

where N is the number of text tokens, and $N = 77$ when using CLIP tokenizer Radford et al. (2021). Here, the control signal is then injected by modifying the attention maps $\boldsymbol{A}_S$ and $\boldsymbol{A}_T$. Note here that $\boldsymbol{A}_S$ encodes the relationship between the feature $\boldsymbol{V}$ and text embedding $\boldsymbol{p}$, while $\boldsymbol{A}_T$ encodes the relationship between pixels among different frames, without any association to the text.

In more detail, the attention maps are modified by directly weakening the attention map outside of the bounding box via multiplying a constant weakening factor, and strengthening the attention map within the box by adding a Gaussian map. Consider now the masks $\boldsymbol{M}_{\mathcal{B}_S} \in \mathbb{R}^{C\times H\times W\times N} \in [0,1]$ and $\boldsymbol{M}_{\mathcal{B}_T} \in \mathbb{R}^{C\times H\times W\times F\times F} \in [0,1]$ for the spatial and temporal layers, respectively, which encodes whether the pixels

involved in the attention maps are either within (1) the control bounding box or not (0). These maps $\boldsymbol{M}_{\mathcal{B}}$ are anchored by box parameters $\boldsymbol{b} \in (b_l, b_t, b_r, b_b) \in \mathbb{R}^4$ representing top, left, right and bottom coordinates. C, H, W stands for the number of channels, the height, and the width respectively. Trailblazer Ma et al. (2024b) then modifies the attention maps as

$$\bar{\boldsymbol{A}}_* = \underbrace{\boldsymbol{A}_* \odot (\lambda_{\mathrm{w}}(1 - \boldsymbol{M}_*) + \boldsymbol{M}_*)}_{\text{weakening outside}} + \underbrace{\lambda_{\mathrm{s}} \boldsymbol{M}_{\mathcal{G}} \odot \boldsymbol{M}_*}_{\text{strengthening inside}}, \tag{2}$$

where subscript $*$ denotes both spatial and temporal, $\odot$ is the element-wise multiplication, λ_{w} is the weakening factor which we set to λ_{w}=0.001, and λ_{s} is the strengthening factor which we set to λ_{s}=0.15,[1] and $\boldsymbol{M}_{\mathcal{G}}$ is a Gaussian map, defined as[2]

$$\boldsymbol{M}_{\mathcal{G}} = \exp\left(-\frac{(\boldsymbol{I}_{coord} - \boldsymbol{c})^2}{2\boldsymbol{\sigma}^2}\right), \tag{3}$$

where I_{coord} is a tensor of image coordinates, $\boldsymbol{c} = \left(\frac{bl+br}{2}, \frac{bt+bb}{2}\right) \in \mathbb{R}^2$ is the center of the box, and standard deviation $\boldsymbol{\sigma} \in \mathbb{R}^2$ is typically set to one-third of the box height h and width w.[3]

While effective as demonstrated in Ma et al. (2024b), the edit in Equation (2) is not differentiable with respect to the box parameters as it contains discrete borders, especially where the mask $\boldsymbol{M}_*$ transitions either from 0 to 1 or vice versa. Moreover, the shape of the edit, as shown in Figure 3, has sharp discontinuities at the edges. While the Gaussian map allows focusing the attention map to the center of the box, the standard deviation σ of the Gaussian is set to a large enough value to ensure that the attention map is spread over the entire box, but this then leads to a truncated shape with clipped edges. Thus, it is non-trivial to optimize the box parameters with respect to any objective.

3.2 Differentiable attention map editing

To address this, we introduce a differentiable attention map editing method that does not rely on replacement. Specifically, we use the Gaussian edit map $\boldsymbol{M}_{\mathcal{G}}$ as a starting point, make its borders smooth so that it is differentiable. We then propose to edit without relying on the binary masks $\boldsymbol{M}_S$ and $\boldsymbol{M}_T$.

Smooth masks. To prevent the discontinuities shown in Figure 3, we smooth out the borders of the Gaussian edit map $\boldsymbol{M}_{\mathcal{G}}$, with 1D smooth step functions, both in the horizontal and vertical directions. Formally, we write

$$\boldsymbol{M}_{\mathcal{B}} = \boldsymbol{M}_{\mathcal{G}} \odot \boldsymbol{M}_x \odot \boldsymbol{M}_y, \tag{4}$$

where $\boldsymbol{M}_x$ and $\boldsymbol{M}_y$ are the smooth step functions defined as

$$\begin{aligned} \boldsymbol{M}_x &= \mathrm{Sig}\left(\frac{I_u - b_l}{\kappa}\right) \odot \mathrm{Sig}\left(\frac{b_r - I_u}{\kappa}\right) \\ \boldsymbol{M}_y &= \mathrm{Sig}\left(\frac{I_v - b_t}{\kappa}\right) \odot \mathrm{Sig}\left(\frac{b_b - I_v}{\kappa}\right) \end{aligned}, \tag{5}$$

where κ controls the strength of the smooth edge transition, and $I_u, I_v \in H \times W$. Formally, strength $\kappa = \lambda_{edge}\sqrt{h^2 + w^2}$, is calculated as a fraction of the bounding box's diagonal length, where h, w are the height and width of the box. In practice, we set $\lambda_{edge} = 0.03$. We show an example of our attention map edit in Figure 3d.

Differentiable editing. It is important to note that differentiability is not guaranteed as long as the discrete mask $\boldsymbol{M}_*$ is utilized . We note, however, our masks $\boldsymbol{M}_{\mathcal{B}}$ are now of similar shape as the binary

[1] Trailblazer Ma et al. (2024b) uses a per-animal hyperparameter setting, but this is impractical for our evaluation scenario with hundreds of animals. We thus use this value to a fixed value that we found empirically and use it for all evaluations, including our method.

[2] In the case of the temporal attention map, there are two centers, one for each frame, and we simply take the maximum value for each pixel for overlapping regions.

[3] The original paper claims one-half, but the official code release uses one-third, which is what we use.

ones, and can thus be used instead of $\boldsymbol{M}_*$ in Equation (2). We thus write:

$$\bar{\boldsymbol{A}}_* = \underbrace{\boldsymbol{A}_* \odot (\lambda_{\text{w}}(1 - \boldsymbol{M}_{\mathcal{B}}) + \boldsymbol{M}_{\mathcal{B}})}_{\text{weakening outside}} + \underbrace{\lambda_{\text{s}} \boldsymbol{M}_{\mathcal{B}}}_{\text{strengthening inside}}, \tag{6}$$

where again, λ_{w} and λ_{s} are the weakening and strengthening factors, respectively, which we set empirically as λ_{w}=0.001 and λ_{s}=0.15 same as in the case of Trailblazer Ma et al. (2024b). While the difference between Equation (6) and Equation (2) is subtle, Equation (6) has no discrete borders, and is thus safe to differentiate with respect to the box parameters.

3.3 Optimizing bounding boxes to align with attention maps

With the attention map editing now being differentiable with respect to the box parameters, we can optimize the bounding box such that the attention map of the next layer is maximized within the box.[4] The difficulty in doing so, however, is how to define what is a good bounding box edit. Our hypothesis is that a good edit would be one that is 'easy' for the neural network to follow, thus being close to what the neural network itself would do. This is because both attention map edits in Section 3.2, are heuristically designed and there is no guarantee that such edit would not derail the video generation process. In fact, as shown earlier in Figure 1, this does happen.

We thus propose to look into how the edited attention maps 'propagate' through the network layers, and optimize the bounding boxes such that this propagation follows the user's intent. Specifically, without loss of generality, let us denote the attention layer that is applied to the edited attention map as f, the 'values' in the typical attention operation as $\boldsymbol{V}$, we can then write

$$\boldsymbol{A}_*^{l+1}, \boldsymbol{V}_*^{l+1} = f(\bar{\boldsymbol{A}}_*^l, \boldsymbol{V}_*^l), \tag{7}$$

where the superscript l denotes the layer index, the subscript $*$ again denotes both spatial and temporal, and $\bar{\boldsymbol{A}}_*^l$ is the edited attention map of layer l. We then look at $\boldsymbol{A}_*^{l+1}$, which now represents how the network is utilizing edited attention, and aim to encourage the attention to be within the user-specified bounding box mask $\boldsymbol{M}_*$.

Inspired by Chen *et.al.* (Chen et al., 2024b), we define the loss to encourage that the sum of all attention values $\sum \boldsymbol{A}_*^{l+1}$ is maximized within the box. This can be expressed in various ways, but as the magnitude of attention may differ from one layer to another, we define it such that the sum of the attention values solely come from within the box, that is $\sum \boldsymbol{A}_*^{l+1} \approx \sum \boldsymbol{A}_*^{l+1} \odot \boldsymbol{M}_*^{l+1}$. We thus write our loss function as

$$\mathcal{L}_{\text{attn}} = \left\| 1 - \frac{\sum \boldsymbol{A}_*^{l+1} \odot \boldsymbol{M}_*^{l+1}}{\sum \boldsymbol{A}_*^{l+1}} \right\|_2^2, \tag{8}$$

Note that this loss then scales between 0 and 1, with 0 indicating the attention map is purely concentrated to be within the bounding box.

Preventing attention from completely ignoring outside the box. Solely relying on the maximization of attention inside the bounding box causes the model to completely ignore the background, resulting in poor generation quality and less alignment to the prompt. Thus, we introduce a balancing loss, which allows our method to attend to the outside region, leading to an effective whole generation. We write

$$\mathcal{L}_{\neg\text{attn}} = \left\| 1 - \frac{\sum \boldsymbol{A}_*^{l+1} \odot \left(1 - \boldsymbol{M}_*^{l+1}\right)}{\sum \boldsymbol{A}_*^{l+1}} \right\|_2^2, \tag{9}$$

which now enforces outside of the box to retain attention.

Regularizing to remain close to user intent. As we wish to preserve as much of the user intent as possible, we regularize to keep the box close to the original one. We write

$$\mathcal{L}_{\text{reg}} = \|\boldsymbol{b} - \boldsymbol{b}_{\text{user}}\|_2^2, \tag{10}$$

[4]Note that this alone causes the model to highly ignore outside the bounding box, which we balance; for ease in explanation, we omit this for now.

where $\boldsymbol{b}$ would be the optimized bounding box, and $\boldsymbol{b}_{\text{user}}$ the original box provided by the user.

Final optimization objective. The final loss is then,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{attn}} + \lambda_{\neg\text{attn}}\mathcal{L}_{\neg\text{attn}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{11}$$

where $\lambda_{\neg\text{attn}}$ and λ_{reg} are the regularization strength, which we empirically set to λ_{reg}=0.1 $\times \sqrt{A}$, where A is the number of pixels in the image and $\lambda_{\neg\text{attn}}$=10.

3.4 Implementation details

We implement our method in PyTorch Paszke et al. (2019) based on the official code of Trailblazer Ma et al. (2024b). We further apply our method from Trailblazer Ma et al. (2024b) to T2V-Turbo Li et al. (2025), a different video generator. For a fair comparison, we use the same video length of 24 frames, and the default setting for all other hyperparameters associated with Trailblazer Ma et al. (2024b). Because of our resource constraints—memory of our GPU cards is 32Gb—we use a resolution of 256 $\times$ 320 and 320 $\times$ 320 for T2V-Turbo Li et al. (2025) and Trailblazer Ma et al. (2024b), respectively. We also use 320 $\times$ 320 resolution for the Peekaboo baseline. We find that λ_{s}=0.3 works well for the T2V-Turbo Li et al. (2025) backbone. Following Ma et al. (2024b), we only edit the spatial attention map. We run 5 optimization iterations for 5 editing steps, leading to 25 iterations in total. We use the Adam optimizer Kingma & Ba (2015) for box adjustments. We will provide code upon acceptance.

4 Experiments

4.1 Experimental setup

Evaluating controllable T2V models is challenging as generated videos lack ground truth and quality metrics only form a proxy to human-perceived quality. It is also non-trivial to generate large and diverse, input control trajectories without violating physics, as the generation outcome is highly sensitive to the input control, as we will demonstrate. For the former, we run a user study alongside human-preference metrics. For the latter, we use control trajectories from animal videos. In the following subsections, we will first discuss the experimental setup, including the dataset and how we create the control trajectories. We also discuss the baselines, the quantitative metrics, and the user study protocol.

Dataset and bounding-box trajectories. As in Ma et al. (2024b), we use the Animal Kingdom dataset Ng et al. (2022) as a reference dataset for actual videos. This dataset contains several wild animals such as cheetah and hippotamus, with corresponding text describing activity, e.g., *the cheetah is running*. This dataset contains 18,744 video clips.

We keep only video clips with a single object using NLTK library Bird et al. (2009), and moving verbs indicating motion. We then use OWL large-scale open vocabulary detector Minderer et al. (2022) to detect object boxes, PySceneDetect Castellano (2014–2024) library to split scenes to extract trajectories, i.e., sequence of bounding boxes. To obtain trajectories that are useful, we drop videos that have less than two frames with the object being detected, or with only two detections across scene cuts. Also, videos that show discontinuous trajectories such as abrupt camera motions or frame cuts are dropped. These are trajectories that have Intersection over Union (IoU) between the consecutive frames being less than 50. The above results in 1,980 videos (157 unique animals) for training and 526 videos (96 unique animals) for testing. We interpolate detected boxes for the missing frames to obtain a complete trajectory over the video. We further filter out trajectories that are of boxes that are too small, i.e., maximum box width and height smaller than 10% of the width and height of the image, and then randomly sample an interval of 24 frames from the trajectory. This results in 377 trajectories from the training set that can be used for method design and validation, and 226 trajectories exclusively for testing.

Baselines. We compare our method against Peekaboo Jain et al. (2024), showcasing the quality improvement of our method. We also compare against Trailblazer Ma et al. (2024b), with its original backbone and we further adapt it to a different recent text-to-video model T2V-Turbo Li et al. (2025). We also compare

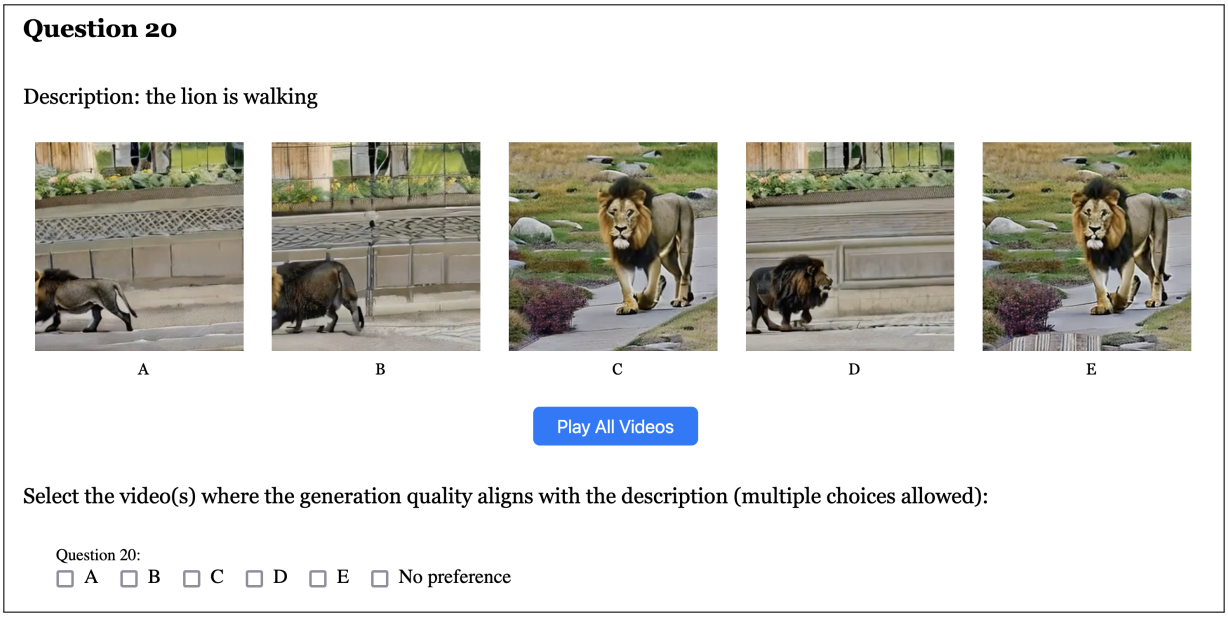


Figure 4: **User study interface** – We show a sample of the user study interface for prompt 'the lion is walking'. Users are asked to select their preference over a set of five video generations, provided in a random order. Users are allowed to select multiple choices or no preference. In this example, only the generation results are shown without user control to isolate quality vs. control. For this question, C and E look preferable, whereas the control is located at the bottom-left where the lion is at A, B, and D. C and E completely ignore user control, yet generate a preferred view of the prompt. As we are interested in controlled generation, we consider both answers to the quality preference and the trajectory faithfulness when evaluating the quality of generations.

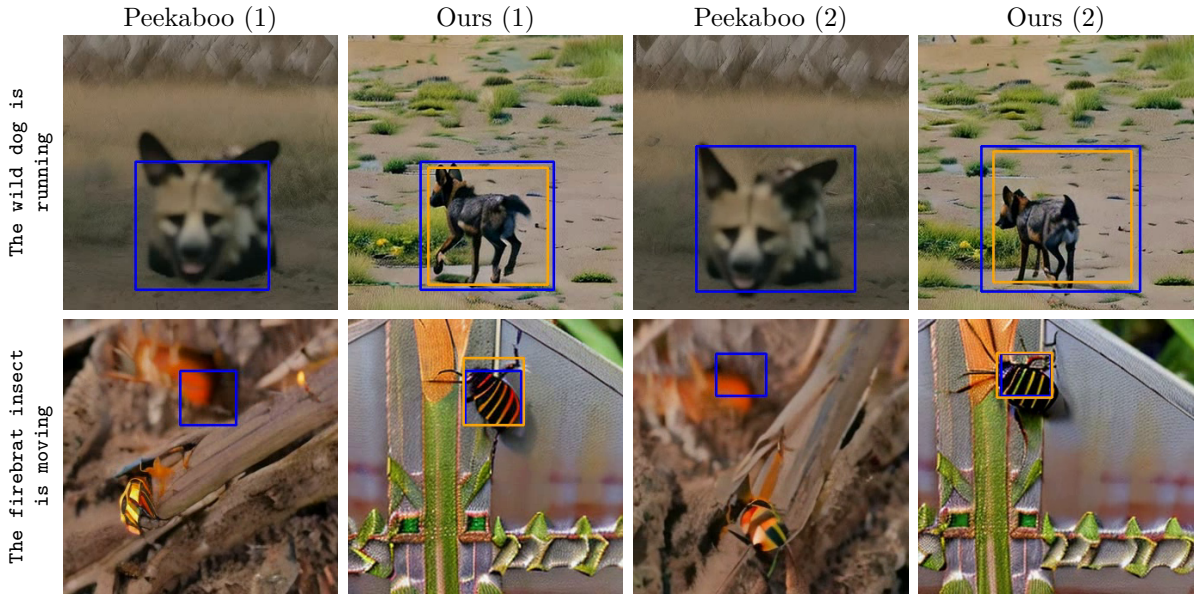


Figure 5: **Qualitative results comparing Peekaboo Jain et al. (2024) and our method.** – Each row shows two representative frames per method (left to right: Peekaboo, Ours). Our method yields better generation quality and improved alignment with the text prompt.

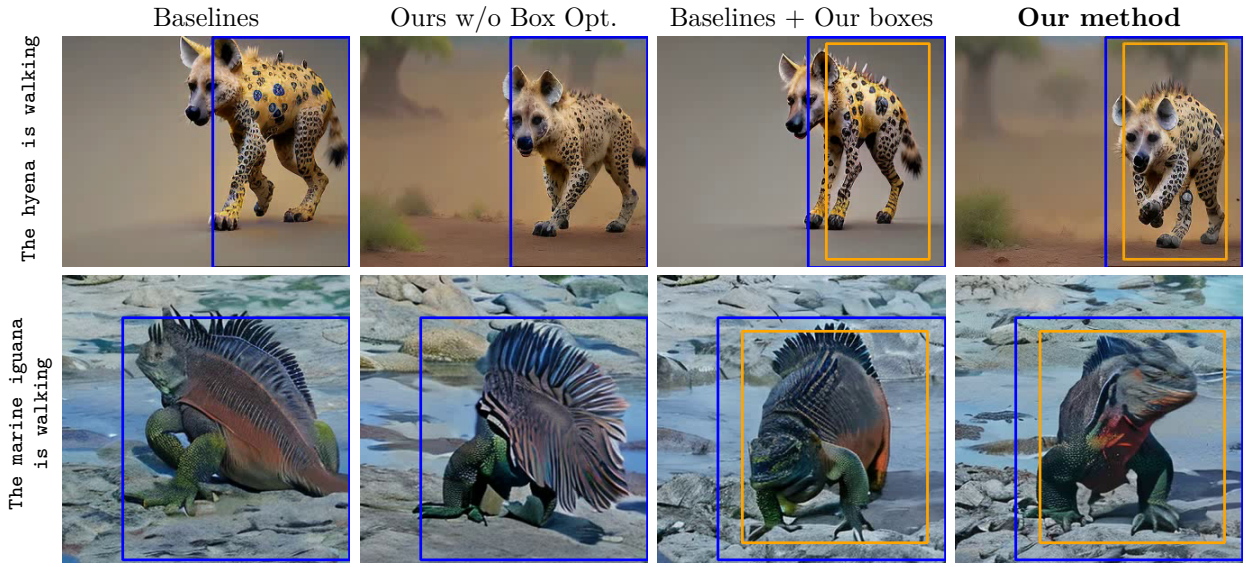


Figure 6: **Qualitative results with T2V-Turbo Li et al. (2025) (top row) and Trailblazer Ma et al. (2024b) (bottom row) backbones.** – We display the original user control in blue and our optimized boxes in orange. As shown, ours provides improved rendering quality and adherence to control and the prompt, shown in the top and bottom row. See video results for best viewing.

our method against different variations of our method, specifically, our method without box optimization, our method without Equation (9), and using our optimized boxes with Trailblazer Ma et al. (2024b).

Quantitative metrics. We quantify the performance of each method using human-preference alignment metrics. We report PickScore Kirstain et al. (2023) and HPS v2 (Human Preference Score) Wu et al. (2023). We also report the mean intersection over union (mIOU) between the detected objects via an open-vocabulary detector Minderer et al. (2022) and the user control, using the animal kingdom dataset as reference.

Metric	PickScore ↑	HPSv2 ↑	mIOU ↑
Trailblazer Ma et al. (2024b)	0.244	0.222	0.37
Our boxes + Trailblazer backbone	0.257	0.223	0.36
Our method w/o Box Opt.	0.243	0.221	0.37
Our method (full)	0.257	0.225	0.37
Peekaboo Jain et al. (2024)	0.149	0.189	0.30
Trailblazer Ma et al. (2024b)	0.175	0.222	0.37
Trailblazer + T2V-Turbo backbone	0.290	0.253	0.41
Our method using T2V-Turbo backbone	0.386	0.263	0.41

Table 1: **Quantitative results with human-preference and control metrics** – Our full method outperforms baseline and demonstrates consistent preference across different architectures, while achieving competitive control.

User study evaluation. We also conduct a user study 20 participants selecting their preference over a set of 5 different methods; see Figure 4. Users were asked to perform two different task: (1) to evaluate the quality of the generations without being provided any bounding-box annotations to solely evaluate quality; and (2) to evaluate the adherence to the bounding box. These were provided as two separate questions and the questions were grouped by tasks. Multiple choices were allowed, including no preference. Each participant was asked to answer 40 questions (20 for T2V-Turbo Li et al. (2025) and 20 for Trailblazer Ma et al. (2024b)) which takes about 20–30 minutes.

To rule out generations that completely failed due to the limitations of the video diffusion model, we chose only those which provide an IoU value of at least 60% for *at least one of the methods* being compared. Candidate questions were then selected randomly and anonymized. A final set with overall quality is used for the study. This random selection was fixed for all participants.

Finally, as the first task ignores the user control completely (following user study design), the preference recorded for this tasks can be irrelevant to the controlled generation task as shown in Figure 4. Thus, we mark as preferred quality only when user selected the video for both tasks. This study was approved by the Institutional Review Board.

4.2 Results

Quantitative results. We report the standard quantitative results in Table 1. Overall, our method demonstrates consistent improvements across different architectures Ma et al. (2024b); Li et al. (2025) (top and bottom row). Also, our method outperforms baselines such as Peekaboo Jain et al. (2024) and Trailblazer Ma et al. (2024b) in terms of human preference scores using PickScore and HPSv2 (bottom row). Additionally, *"Our boxes + Trailblazer Ma et al. (2024b) "* validates the benefits of applying our adjusted boxes (top row). In terms of control accuracy, the results show that our method matches the same performance as the baselines. Though the benefit of adjustments is well highlighted in the user studies, this results reveals that our adjustments do not degrade control.

Qualitative results. We show qualitative examples of video frames generated by our method compared to baselines in Figure 5 and Figure 6. Notably, our method perform significantly better than Peekaboo Jain et al. (2024) in terms of generation quality and adherence to control in Figure 5. Particularly, our method produces better quality, as shown in the *dog* example (top row) and adherence to control, as shown in the *firebrat insect* example (bottom row).

Also when compared with T2V-Turbo and Trailblazer backbones, our method, courtesy of adjustments, produces better generation that adhere with the prompt in Figure 6. This is especially visible in the second row, where generation follows motion in the prompt *"The marine iguana is walking"*. Interestingly,

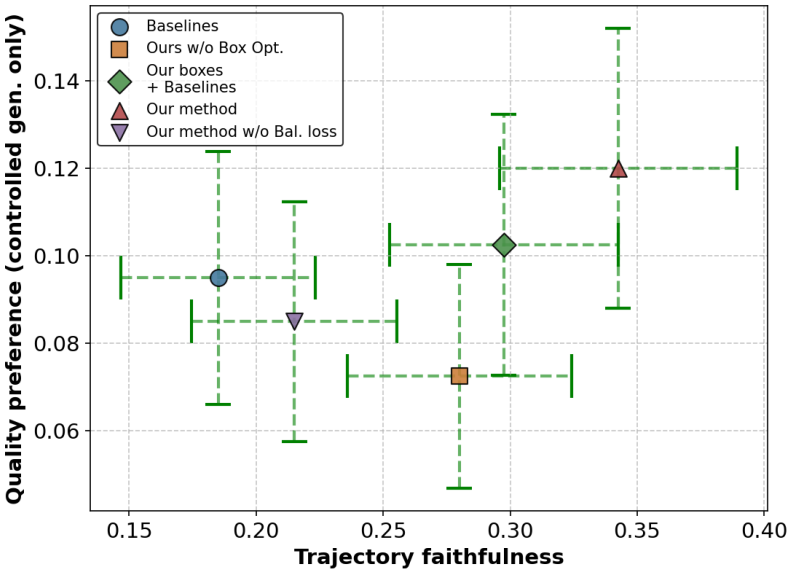


Figure 7: **User study results** – We report the user preference for trajectory faithfulness on the x-axis, and quality of generation (for generations that adhere to control) on the y-axis. The 95% confidence intervals are drawn as green dashed lines. Our method significantly outperforms Trailblazer Ma et al. (2024b) with both backbones.

applying our adjusted bounding boxes also provides better prompt following for the baseline method. Though this benefit is not always transferable, as the baseline still uses a discrete editing strategy. Best performance with adjustment is achieved with our method. Finally, in Figure 8 and Figure 9, we show additional qualitative results of our method compared to baselines.

User evaluation. We report user study results in Figure 7. As shown, our method is the most preferred by a large margin. It is also worth noting that using our adjusted boxes also benefits Trailblazer Ma et al. (2024b), as shown by the slight increase in preference. We hypothesize that this is because our boxes are optimized for the same underlying model. This further demonstrates that small adjustments matter, not only for the method that we have introduced, but also for the base model. Finally, Ours w/o balancing loss (i.e., Eqn. (9) being left out) show that taking the overall generation into account is effective. Simply optimizing the bounding box, without considering the whole generation results in a middle ground between the baseline and the non-optimized version. Only when both are enabled, we can achieve a significant performance increase.

5 Conclusion

In this work, we demonstrated that small adjustments to user-provided bounding boxes can lead to significant improvements in controlled video generation. By optimizing the bounding boxes to better align with the internal attention maps of video diffusion models while maintaining proximity to user inputs, we achieved both higher quality generations and better adherence to control signals. Through extensive experiments and user studies, we validated that our simple yet effective approach outperforms existing methods for controlled video generation. We believe our findings open up new possibilities for improving user control in text-to-video generation by considering how control signals can be optimized to work better with pretrained models.

Limitations and future work A limitation of our method is that the quality of generated images can be bound by the underlying video model, as shown in Figure 10. With the rapid progress in this area, we believe this limitation will be alleviated naturally. While in theory our method can be applied to other forms of

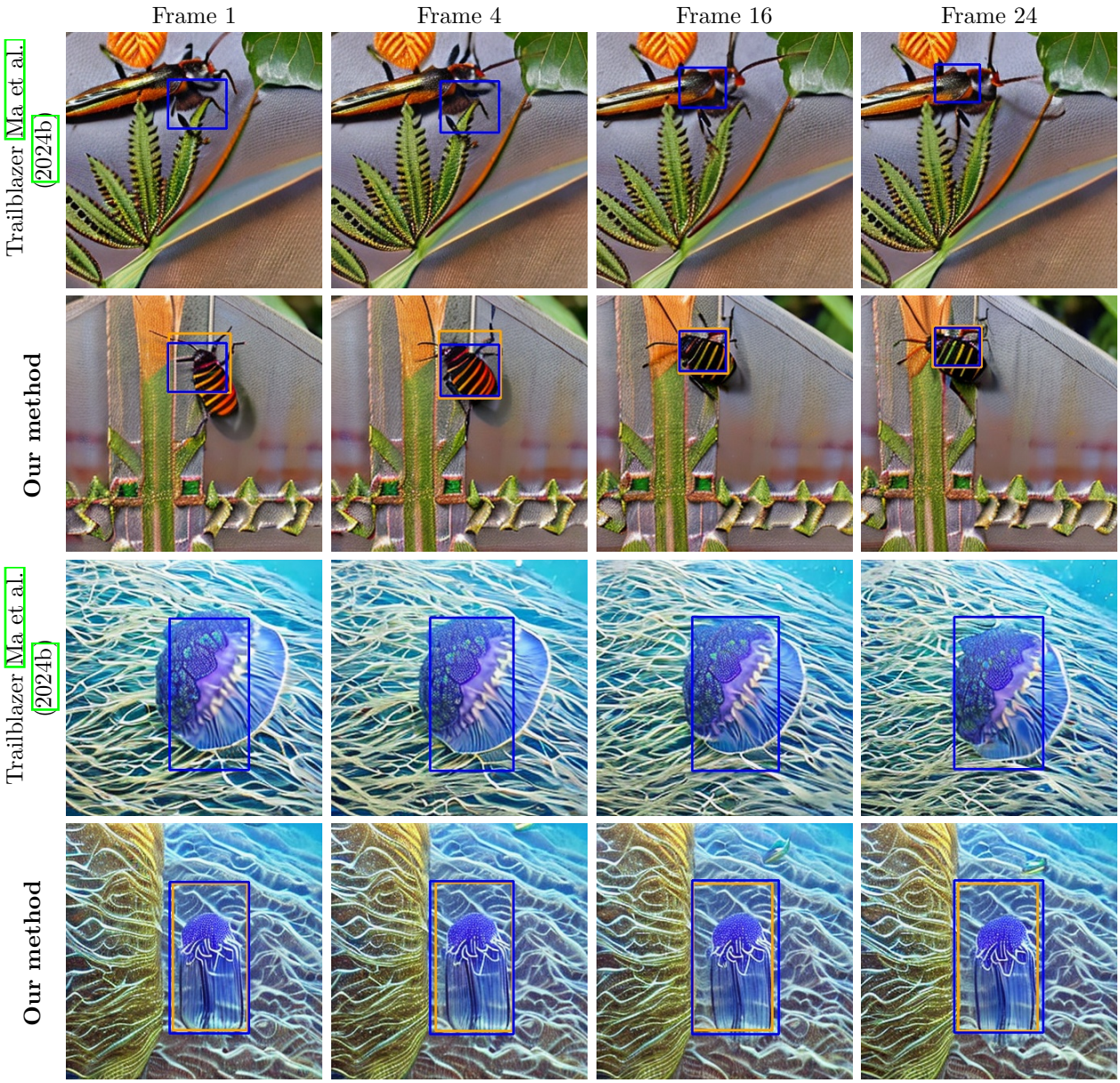


Figure 8: **Additional qualitative results with Trailblazer Ma et al. (2024b)**– We show additional examples of our edits. As shown, ours is more consistent with user intent and of higher quality. The prompt for the top 2 rows is *"The firebrat insect is moving"*, while the prompt for the bottom 2 rows is *"The jellyfish is swimming"*. Video results are available in the supplementary local HTML page.

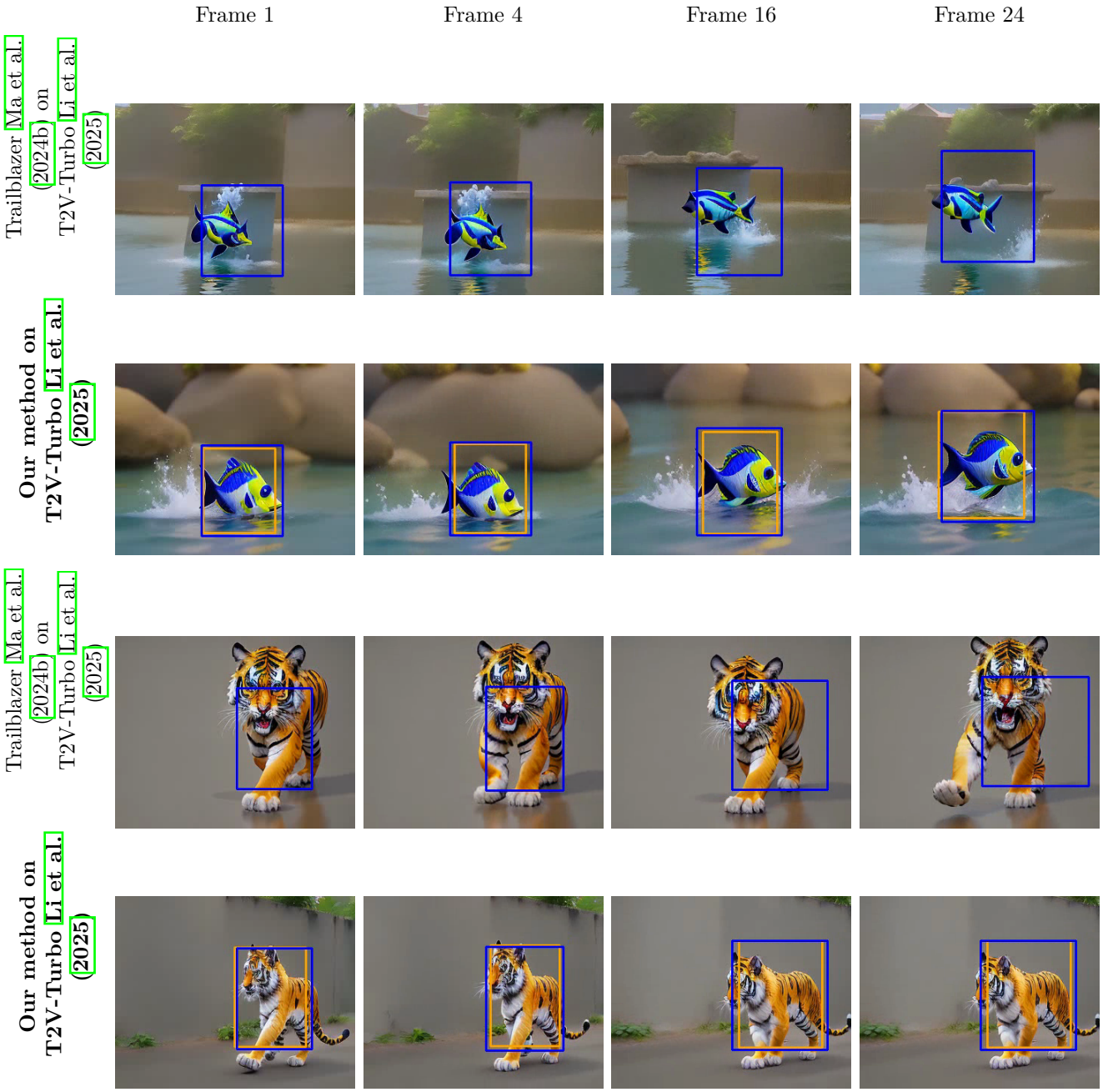
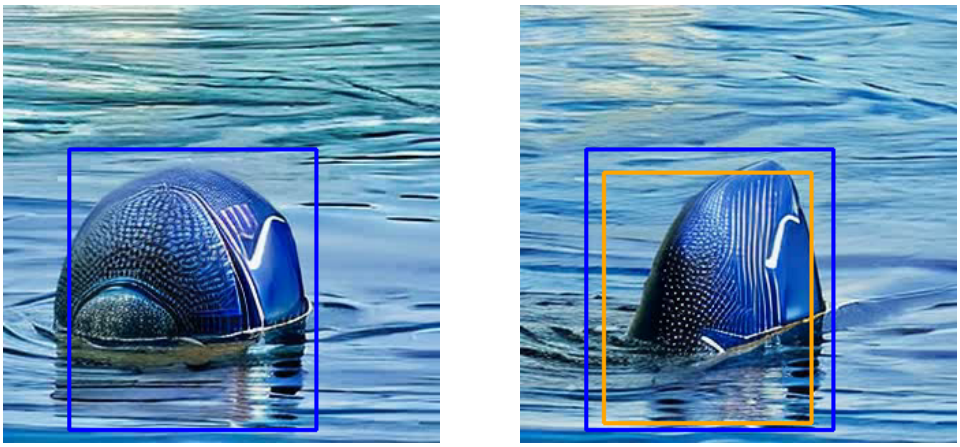


Figure 9: **Additional qualitative results with T2V-Turbo Li et al. (2025) backbone** – We show additional examples of our edits. Ours show more consistency with user intent and is of higher quality within the box. The prompt for the top 2 rows is *"The surgeonfish is swimming"*, while the prompt for the bottom 2 rows is *"The tiger is walking"*. Video results are also available in the supplementary local HTML page.



(a) Trailblazer Ma et al. (2024b) (b) Our method

Figure 10: **Failure cases** – We show example failure cases when the video model fails to generate the desired output for the prompt `the orca is swimming`. Here, regardless of the bounding box input, the model is unable to generate the content as it has no knowledge of the target prompt. We hypothesize that more advanced video models, which are being made increasingly available Research (2023); Wan et al. (2025), will alleviate this problem.

control, we have only validated our idea to bounding boxes. Investigating user-control-adjustments to other control methods would be interesting future work. Finally, our method requires partial back-propagation during optimization, thus slows down generation. With the T2V-Turbo Li et al. (2025) backbone, our generations take one minute and 37 seconds, while original non-controlled generation takes 47 seconds (both on a NVidia RTX A6000). While our edits bring significant enhancements, optimizing this would also be an interesting extension. Finally, our main novelty demonstrates that small adjustments matter, and we therefore wish to draw the attention of the community to the sensitivity of control signals and hope that our work will serve as a foundational ground.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

Steven Bird, Edward Loper, and Ewan Klein. Natural language toolkit, 2009. URL https://www.nltk.org/.

Brandon Castellano. Pyscenedetect: Video scene cut detection and analysis tool, 2014–2024. URL https://www.scenedetect.com/. Version 0.6.4.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning*, 2023.

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.

Changgu Chen, Junwei Shu, Lianggangxu Chen, Gaoqi He, Changbo Wang, and Yang Li. Motion-zero: Zero-shot moving object control framework for diffusion-based video generation. In *The Association for the Advancement of Artificial Intelligence*, 2025.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024a.

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*, pp. 5343–5353, 2024b.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations (ICLR)*, 2023.

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*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022b.

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8079–8088, 2024.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023.

Cheng Lei, Jiayu Zhang, Yue Ma, Xinyu Wang, Long Chen, Liang Tang, Yiqiang Yan, Fei Su, and Zhicheng Zhao. Ditraj: training-free trajectory control for video diffusion transformer. *arXiv preprint arXiv:2509.21839*, 2025.

Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. In *International Conference on Learning Representations (ICLR)*, 2025.

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Ge Ya Luo, ZhiHao Luo, Anthony Gosselin, Alexia Jolicoeur-Martineau, and Christopher Pal. Ctrl-v: Higher fidelity autonomous vehicle video generation with bounding-box controlled object motion. *Transactions on Machine Learning Research*, 2025.

Wan-Duo Kurt Ma, Avisek Lahiri, John P Lewis, Thomas Leung, and W Bastiaan Kleijn. Directed diffusion: Direct control of object placement through attention guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4098–4106, 2024a.

Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024b.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pp. 728–755. Springer, 2022.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19023–19034, 2022.

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. In *International Conference on Machine Learning*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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*, pp. 2085–2094, 2021.

Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. *arXiv preprint arXiv:2406.16863*, 2024.

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*, pp. 8748–8763. PMLR, 2021.

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.

Runway Research. Genmo: Text-to-video generation with diffusion models, 2023. URL https://runwayml.com/. Runway Research Genmo, accessed 2024-11-14.

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*, pp. 10684–10695, 2022.

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. In *International Conference on Learning Representations (ICLR)*, 2023.

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*, pp. 1921–1930, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. In *International Conference on Machine Learning*, 2024.

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*, 36:7594–7611, 2023.

Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *International Conference on Learning Representations (ICLR)*, 2020.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv*, 2023.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7452–7461, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22490–22499, 2023.