# AnyI2V: Animating Any Conditional Image with Motion Control

Ziye Li[1] Hao Luo[2,3] Xincheng Shuai[1] Henghui Ding[1]✉
[1]Fudan University [2]DAMO Academy, Alibaba group [3]Hupan Lab
https://henghuiding.com/AnyI2V/

# Appendix

## A.1. More Implementation Details

In this section, we provide a more detailed explanation of our implementation.

**Details of Optimization.** During the latent code optimization stage, we enable fp16 mode to accelerate the process and employ the AdamW [5] optimizer with a learning rate of 0.01.

**Details of DDIM Inversion.** We integrate our method into three frameworks: AnimateDiff [4], Lavie [9], and VideoCrafter2 [1]. AnimateDiff adopts a training strategy where only the newly added temporal attention layer is partially trained, while other modules remain fixed. This approach keeps the base model intact and focuses training on temporal attention modules. Consequently, during the DDIM inversion phase, we disable the temporal attention layer when processing a single image, maintaining consistency with the backbone model design.

On the other hand, both Lavie and VideoCrafter2 adopt a training strategy of joint image-video training, thereby granting them the ability to generate single frames as well. Lavie additionally employs a temporal attention layer, whereas VideoCrafter2 integrates both a temporal attention layer and a temporal convolution layer. When processing a single frame, we disable temporal modules to align the inference process with typical image-generation pipeline.

**Details of loss function.** The target loss function of our method is defined as:

$$\mathcal{L}_j^i = \left\| M_1^i \odot M_j^i \odot \left( F_j[\mathcal{B}_j^i] - \mathrm{SG}\big(F_1[\mathcal{B}_1^i]\big) \right) \right\|_2^2,$$

where $F_1[\mathcal{B}_1^i]$ is a feature that is unrelated to $z_t$. You might wonder why we still employ the stop-gradient operation. The reason is that when optimizing the latent, we utilize the query across different layers. Without stopping the gradient, computing the loss at a later layer would backpropagate gradients through the earlier query, since it remains part of the computational graph.

✉ Henghui Ding (henghui.ding@gmail.com) is the corresponding author with the Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China.

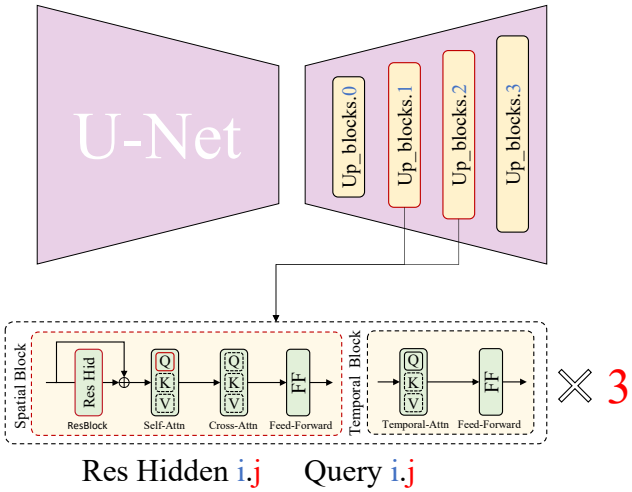


Figure 1. The structure of the video diffusion model.

**Details of Model Structure.** The Fig. 1 shows the explicit model structure of our diffusion model. The decoder consists of four submodules operating at different resolutions. Each submodule contains three spatial blocks and three temporal blocks, and these features are indexed in the format `i.j`, corresponding to the respective `Up_blocks.i` and `Spatial_blocks.j`.

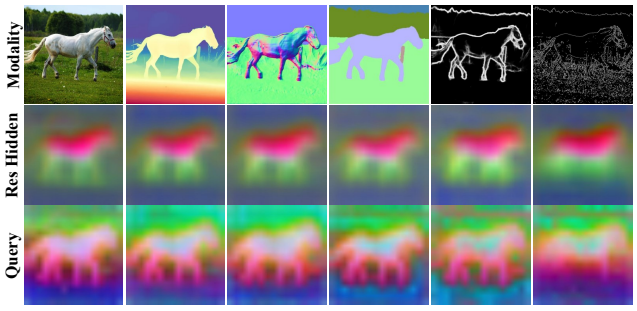


Figure 2. The PCA visualization of residual hidden state and query extracted from different modality images.

**Understanding Images from Different Modalities.** AnyI2V can understand unnatural images because the backbone is capable of capturing similar semantic representations across different modalities. We visualize

the features using PCA by stacking them as a single group. The corresponding result can be seen in Fig. 2

**Controlling multiple objects.** AnyI2V can control multiple objects since our proposed semantic mask can effectively generate masks for each object in bounding box, reducing the impact of overlapping during movement which can be seen in Fig. 3.

Figure 3. Controlling object movement of multiple objects.

**Fidelity Metric for Structure Control.** In the table below (Tab. 1), we measure the structural fidelity of the first frame using the DINO score, by extracting its structural features. AnyI2V underperforms ControlNet in structural control due to its zero-shot nature. For further discussion on structure control, please refer to A.2.

| Structure Similarity | Depth | HED | Canny | Normal | Seg |
|---|---|---|---|---|---|
| AnyI2V (ours) | 0.759 | 0.824 | 0.833 | 0.828 | 0.871 |
| ControlNet | 0.779 | 0.899 | 0.925 | 0.871 | 0.907 |

Table 1. Structural Fidelity Comparison with ControlNet.

## A.2. Further Explanation For Limitations

In the main text, we mentioned some limitations of AnyI2V. In this chapter, we will further explain them by showing some failure cases of our method.

**Controlling Very Large Motion.** Since AnyI2V adopts a latent optimization strategy for motion control, it may struggle with alignment when the target bounding box undergoes significant displacement. As shown in Fig. 4, the generated motion follows the trajectory when the motion amplitude is small to moderate. However, when the motion becomes excessively large, like in SynFMC [7] or other video datasets [2, 3], our method encounters difficulties in maintaining control. Similarly, the training-based DragNUWA [10] also struggles with very large motion, indicating that this remains a common challenge that needs to be addressed.

**Ambiguous Occlusion.** While AnyI2V can handle various spatial conditions, it struggles to determine the correct depth order when overlaps occur, as shown in Fig. 5. Although AnyControl [8] has addressed this issue to a certain extent, resolving it in a training-free way remains an open challenge.

**Reference Frame Control.** Our method controls the first frame by injecting features and regulates motion by optimizing the latent. Since both feature injection and latent optimization occur in the early stages of the diffusion denoising process, the control over the reference image is not as precise as ControlNet [11].


Figure 4. The control results of different motion amplitude.

Figure 5. The control results when control conditions overlap.

To analyze the degree of control over the reference frame, we conduct an experiment to examine the impact of the time step used for latent optimization. Fig. 6 presents the results under different optimization steps. We observe that as the number of optimization steps increases, the first frame progressively aligns more closely with the structure of the conditional image. However, excessive feature injection and latent optimization leads to noticeable degradation in the generated results.

Therefore, to balance output quality and computational efficiency, we optimize the latent only when $t' \leq 5$.

## A.3. Further Explanation of PCA Features

In the main text, we employ PCA-based dimensionality reduction when aligning the query and state that "lower-ranked components exhibit lower temporal consistency".

To further clarify the motivation behind using PCA, we visualize the lower-ranked components. The results in Fig. 7 illustrate PCA components 98–100, which exhibit low temporal consistency and an ambiguous spatial layout.

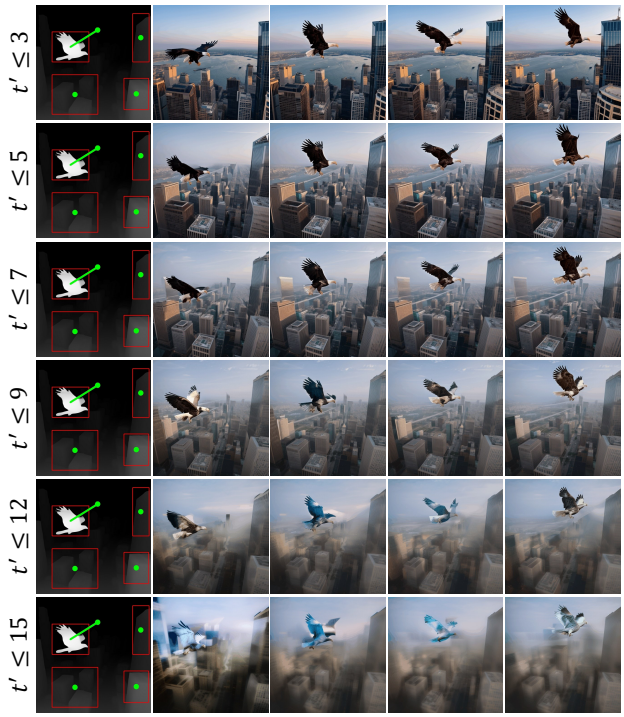


Figure 6. The results of different optimization steps.


Figure 7. Visualization of lower-ranked PCA components.

## A.4. Additional Visual Results

**Extended Visual Results on AnimateDiff.** Fig. 8 presents additional examples of AnyI2V applied to the AnimateDiff [4] backbone, showcasing the diversity and effectiveness of our method.

**Extended Visual Results on Lavie and VideoCrafter.** Fig. 9 illustrates further examples of AnyI2V using the Lavie [9] and VideoCrafter2 [1] backbones, demonstrating its adaptability across different frameworks.

**Examples of Camera Motion Control.** AnyI2V primarily focuses on controlling the motion of objects. However, when an object's attributes remain static, the method can also achieve camera motion effects, as shown in Figure 10. Nonetheless, the camera control capability is currently limited to simple trajectories, highlighting an area for further exploration.

**Examples of Visual Editing.** AnyI2V enables visual editing by utilizing different text prompts [6]. Figure 11 showcases a case where the input structure is a horse. By modifying the text prompt, the generated results maintain a natural structure and appearance, benefiting significantly from the automatically generated semantic mask.

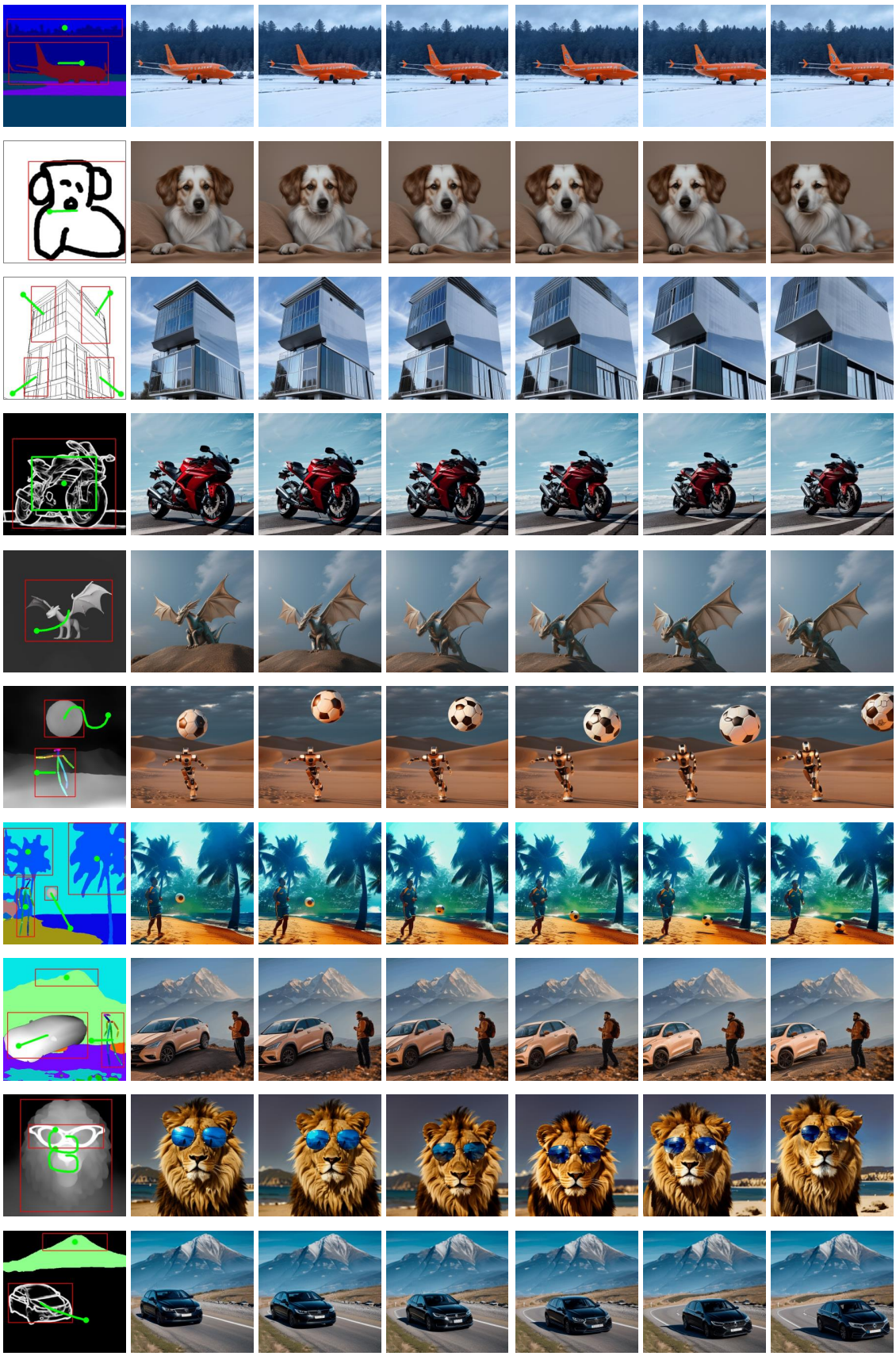

Figure 8. More cases of AnyI2V on the baseline of AnimateDiff [4]


Figure 9. The results on Lavie [9] and VideoCrafter2 [1].

Figure 10. The camera control results of AnyI2V.

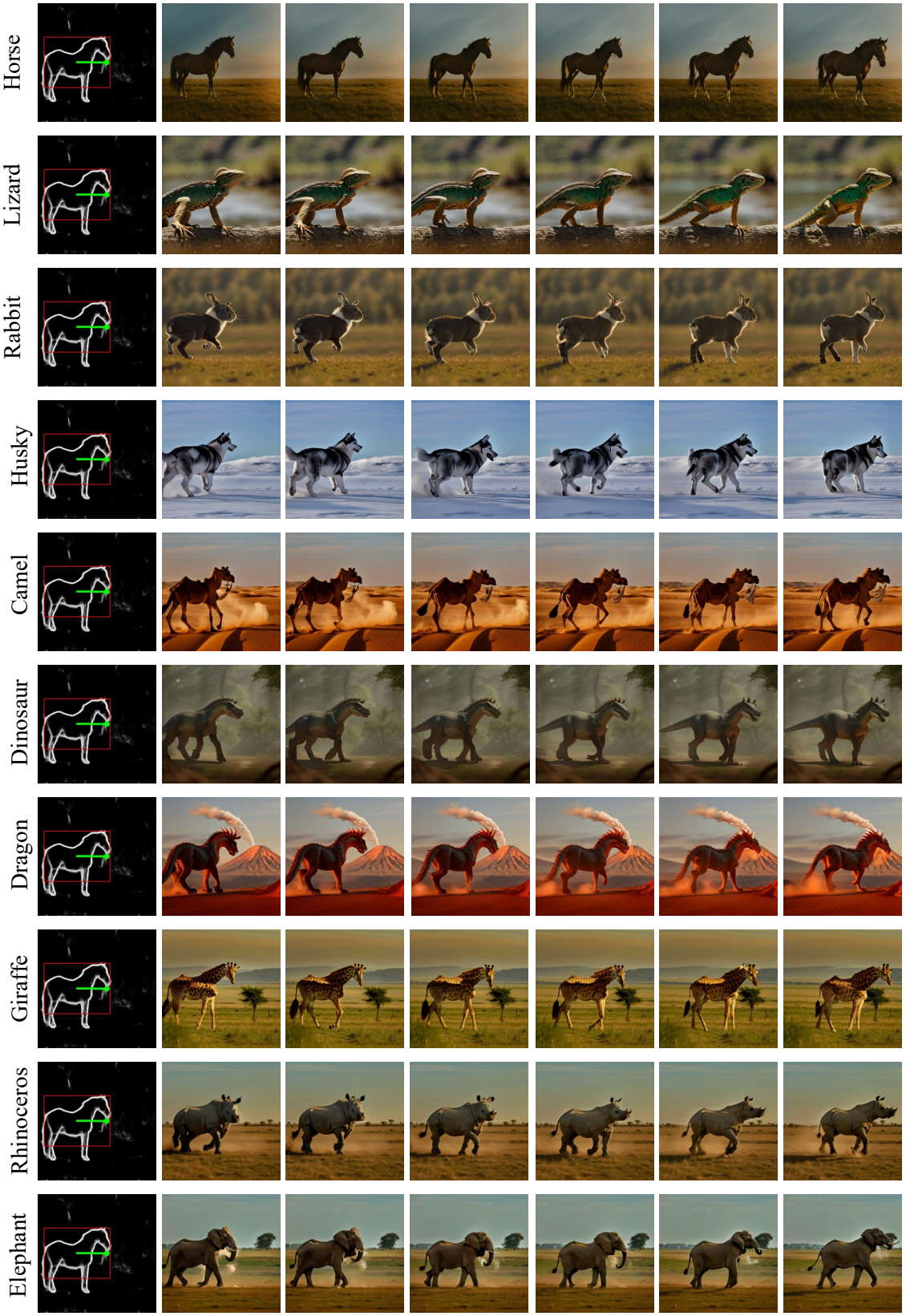


Figure 11. The results of same conditions with different prompts.

## References


[1] 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*, pages 7310–7320, 2024. 1, 3, 5

[2] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 2

[3] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023. 2

[4] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 3, 4

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[6] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024. 3

[7] Xincheng Shuai, Henghui Ding, Zhenyuan Qin, Hao Luo, Xingjun Ma, and Dacheng Tao. Free-form motion control: Controlling the 6d poses of camera and objects in video generation. In *ICCV*, 2025. 2

[8] Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: create your artwork with versatile control on text-to-image generation. In *European Conference on Computer Vision*, pages 92–109. Springer, 2024. 2

[9] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20, 2024. 1, 3, 5

[10] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2

[11] 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*, pages 3836–3847, 2023. 2