Supplementary Materials: Zero-Shot Controllable Image-to-Video Animation via Motion Decomposition

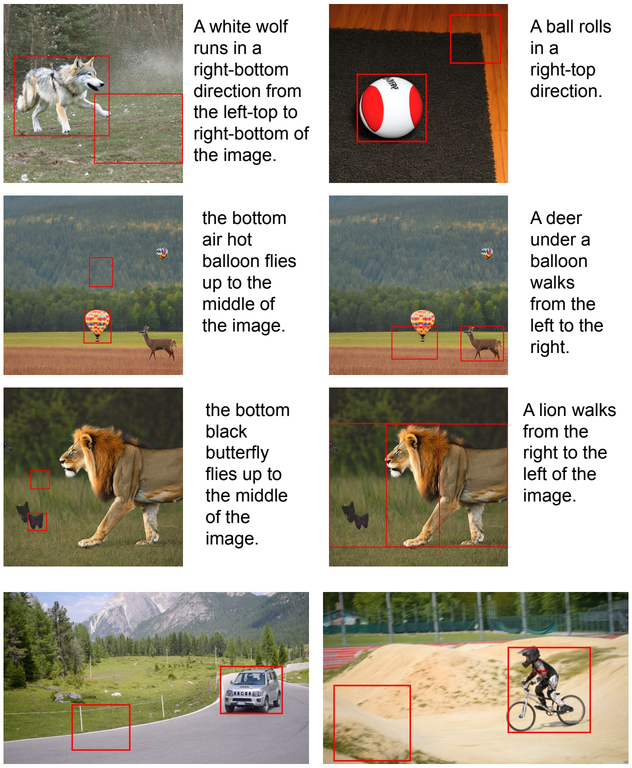
Anonymous Authors

In this Appendix, we present the following:

- Additional details on the proposed benchmark for controllable image-to-video generation evaluation (Sec. 1).
- Attempt and observation on one-shot fine-tuning for better object consistency (Sec. 2).
- Additional qualitative results generated by the proposed IVA^0 and baseline models (Sec. 3).

1 CONTROLLABLE I2V BENCHMARK BUILDING

Data Collection & Annotation: We gather 20 real-world images from public tracking/segmentation datasets [2] alongside 80 synthetic images created using the StableDiffusion-2.0 model [3]. For the synthetic images, we craft specific text prompts to guide the image generation process with StableDiffusion. The images are selected to include either single or multiple objects, ensuring there's a clear, unoccupied space suitable for demonstrating noticeable object motion. To streamline the annotation process, we manually label only the essential layout boxes for each image. These boxes are then interpolated to simulate the effect of user interaction, which aids in generating a sequence of layouts during inference. Furthermore, for each image and its corresponding layout, we create detailed motion-related captions. These captions serve to train baseline models [7] that do not support layout input directly. We display the data—consisting of images, key boxes, and captions—in the format illustrated in Fig. 1.



A white wolf runs in a right-bottom direction from the left-top to right-bottom of the image.

A ball rolls in a right-top direction.

the bottom air hot balloon flies up to the middle of the image.

A deer under a balloon walks from the left to the right.

the bottom black butterfly flies up to the middle of the image.

A lion walks from the right to the left of the image.

A gray car moves on the road, from the right to the left bottom of the image.

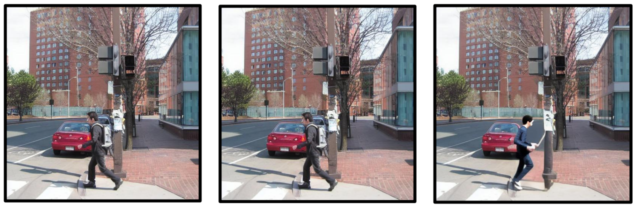
A man rides a bike from the right to the left of the image.

Figure 1: Data examples. We collect both real and generated images for our controllable Image-to-Video benchmark. The red boxes are manually annotated key boxes. We write motion-related captions for each image-layout pair for our baseline method. Best viewed in color.

2 MORE EXPERIMENTS & ANALYSIS

One-shot Fine-tuning for Object Consistency. With the IVA^0 pipeline described in the main paper, we can animate objects in a given image and let them follow user-provided layouts. However, as shown in Fig. 2, we find that although the object appearances are consistent across frames with the help of the motion module, they do not exactly match the object in the given image. One possible reason for this is that we only use CLIP image embedding to control the object appearance, while CLIP can capture high-level semantic features, *e.g.*color and category, but lacks low-level features like shape. Such difference is further amplified when we apply IVA^0 to the real-world images due to the potential domain gap present in the pre-training text-to-image model [1]. To quantify this, we computed the SSIM↑ between ground-truth objects and the reconstructed images. The results indicate that inpainted objects share only 22% SSIM↑ score with the GT, highlighting an unavoidable loss of low-level structural details.

To eliminate this gap, we are motivated by recent one/few-shot tuning text-to-video generation work [5, 6] to try to extend our Img2VidAnim-Zero to a one-shot learning setting. Specifically, we adopt the parameter-efficient fine-tuning strategy, where we freeze the CLIP encoder and the autoencoder in the pipeline, while only updating the CLIP image embedding and the value projection



Original Image

VAE Reconstructed

Zero-shot Inpainted

Figure 2: Comparison among the original image, VAE reconstructed image, and zero-shot inpainted image. We observe the inconsistency issue in the inpainted image due to the lack of low-level features.

inside the diffusion U-Net. In this case, we force the model to use the CLIP-initialized image feature to inpaint the masked initial image x_1 as closely as possible with the standard MSE loss over predicted and GT noises for better object consistency. However, our results show that such one-shot learning is not helpful for object consistency due to worse generation quality. We assume more low-level image features from visual models like VGG [4] are needed for this inconsistency issue. We leave this part for deeper future studies.

Multi-object Animation. To validate the performance in the multi-object cases, we conduct quantitative analysis. For vIoU@0.5 ↑, the multi-object (50 samples) animation scores 86.0% v.s. 86.4% on the single-object (50 samples) animation task. For FVD↓, multi-object animation achieves 1268, only marginally inferior to 1227 of the single-object animation. This confirms that IVA^0 shows competitive performance across multiple and single-object animations.

3 MORE QUALITATIVE RESULTS

In this section, we provide more qualitative results, Fig. 3 to Fig. 8 showcase additional generated video examples from our methods.

REFERENCES

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

[2] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 724–732.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[4] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[5] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. 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*. 7623–7633.

[6] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. 2023. LAMP: Learn A Motion Pattern for Few-Shot-Based Video Generation. *arXiv preprint arXiv:2310.10769* (2023).

[7] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. 2023. DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors. *arXiv preprint arXiv:2310.12190* (2023).

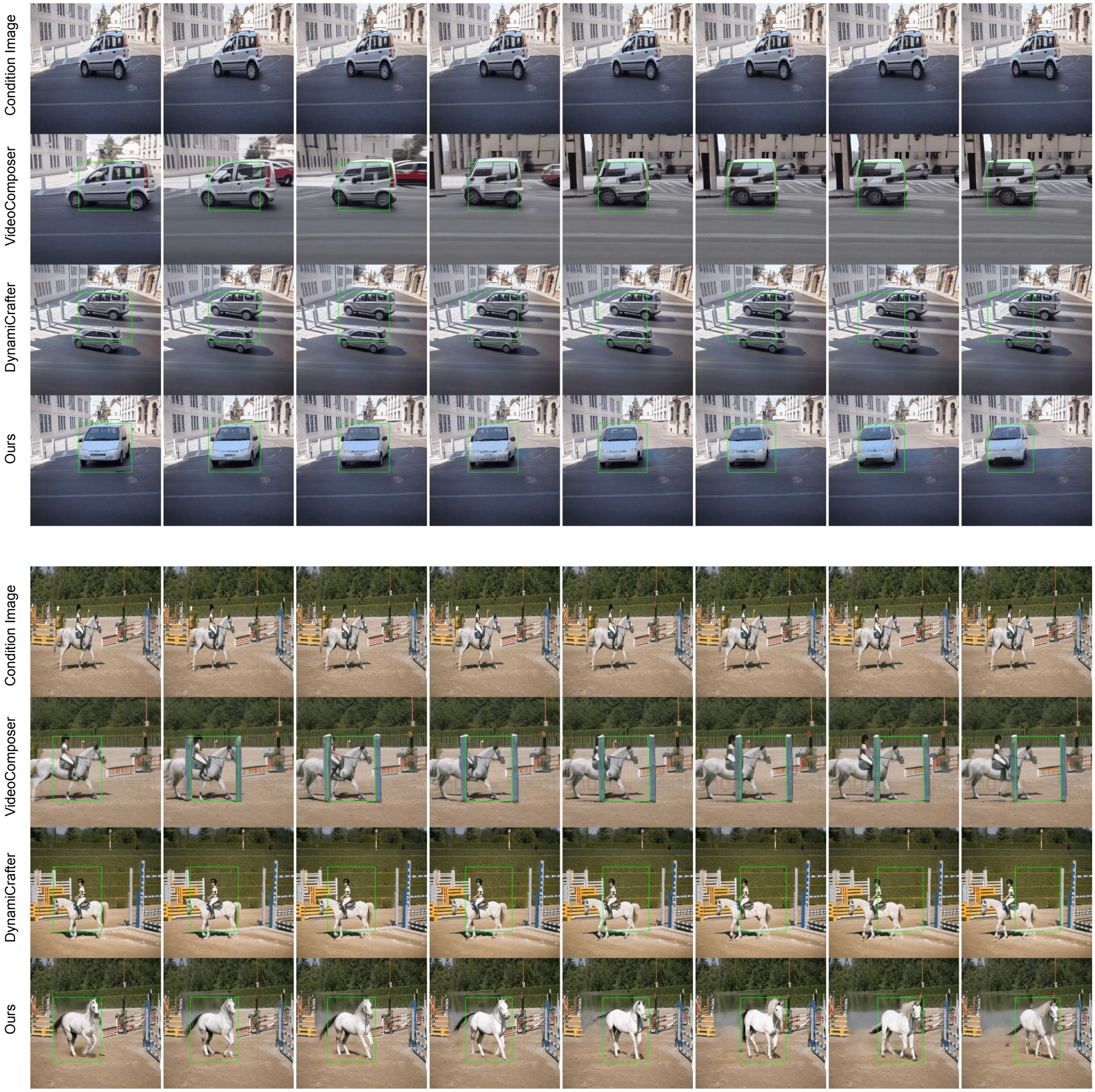


Figure 3: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.

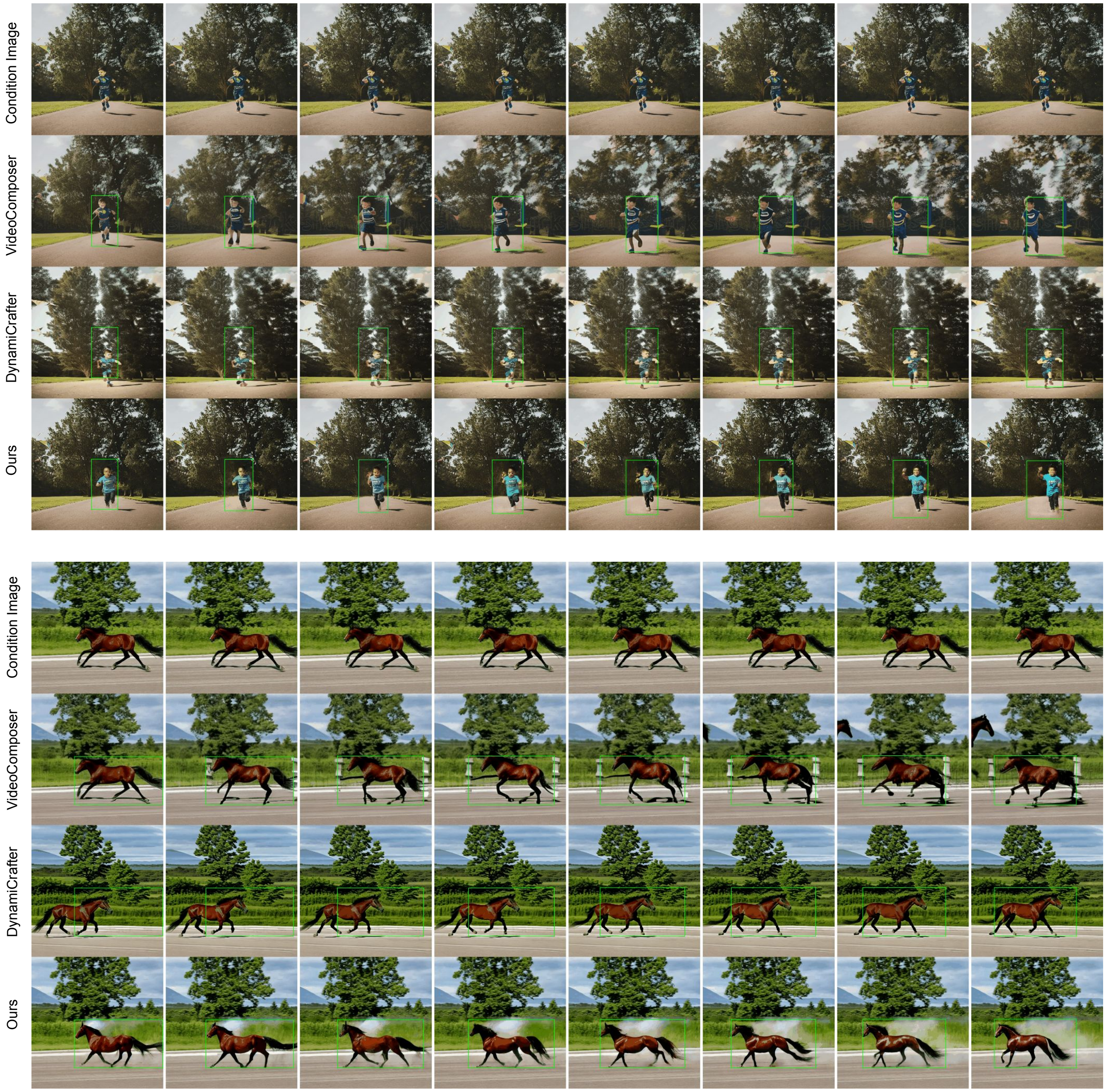


Figure 4: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.

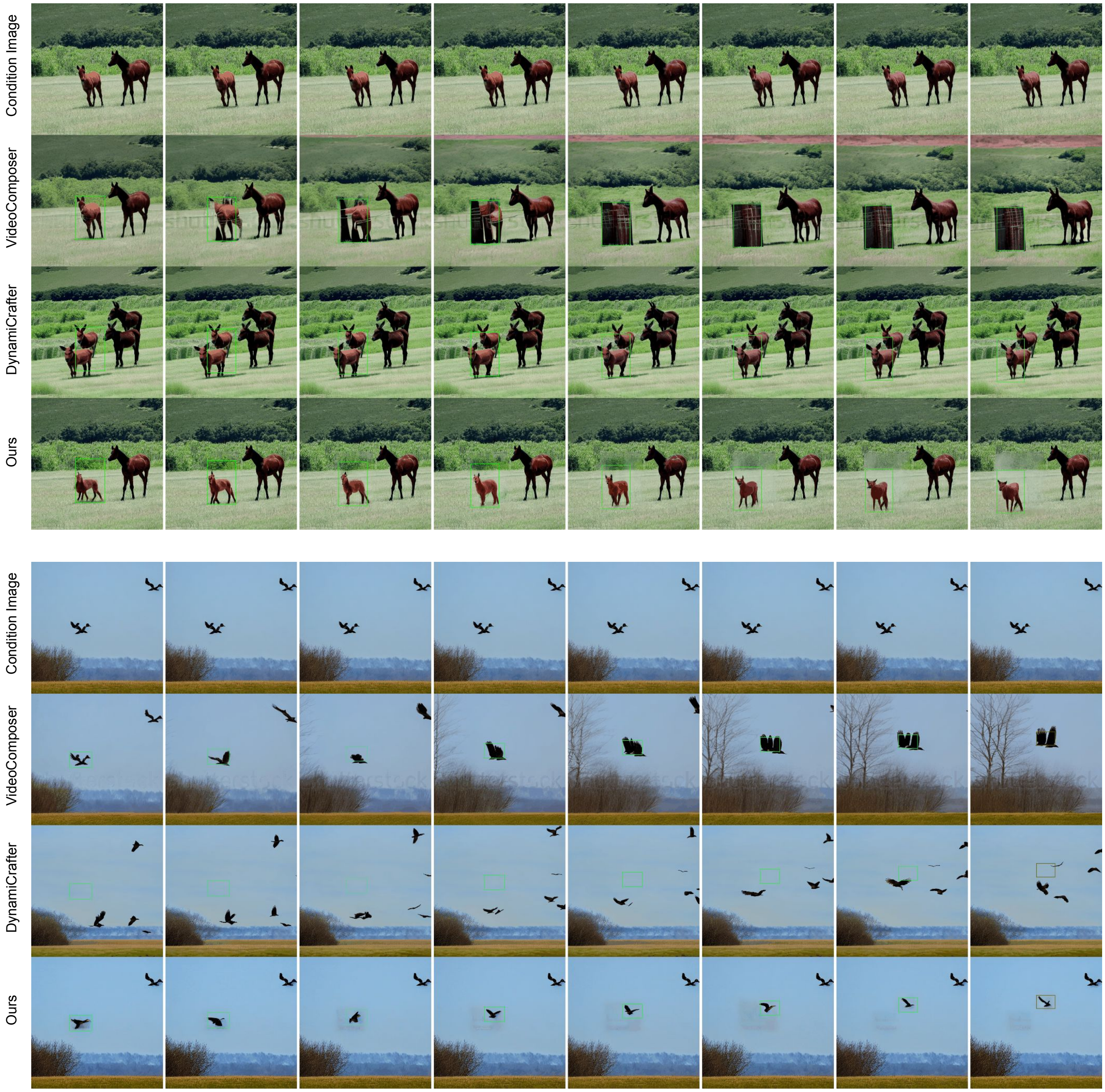


Figure 5: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.

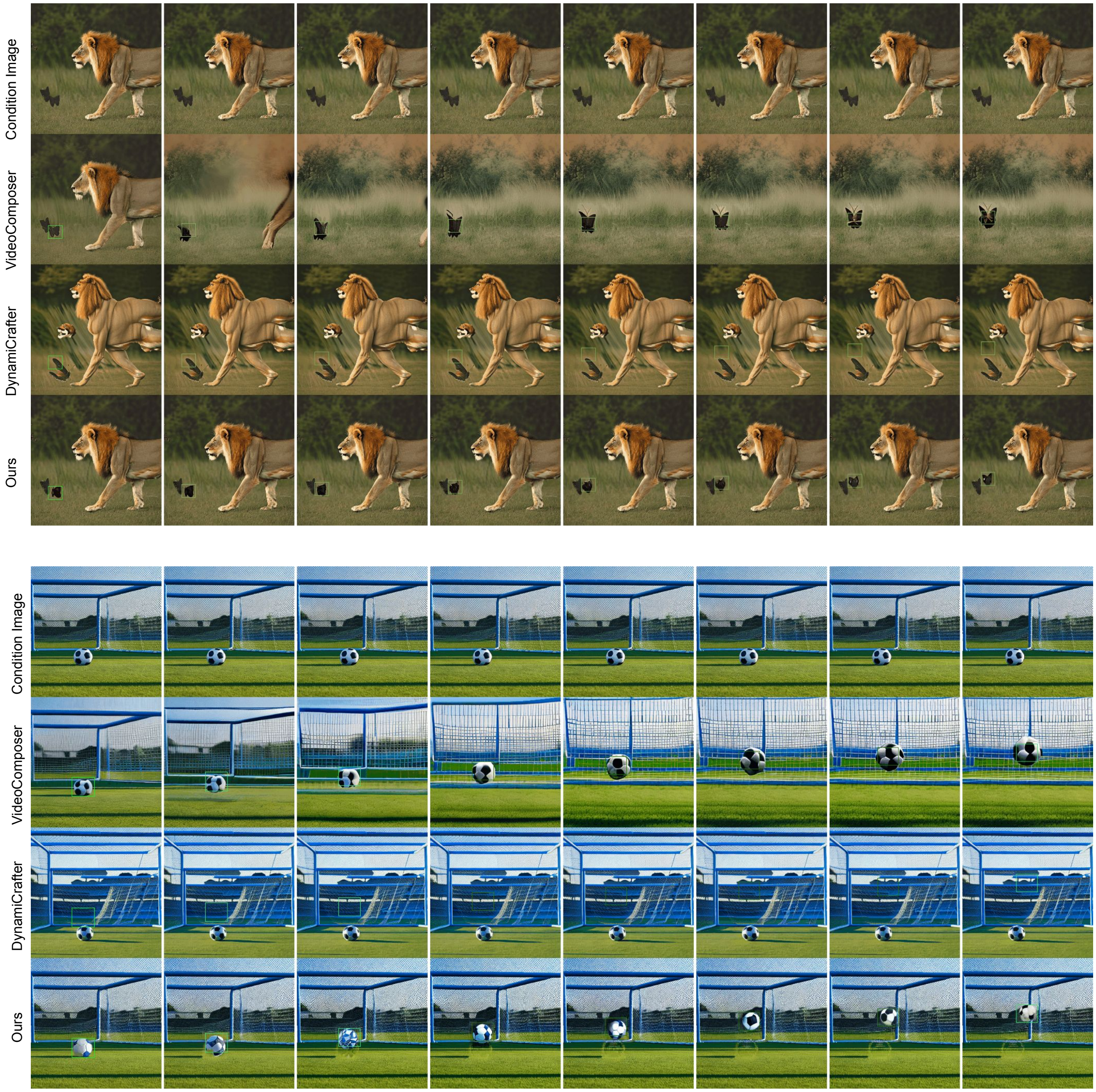


Figure 6: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.

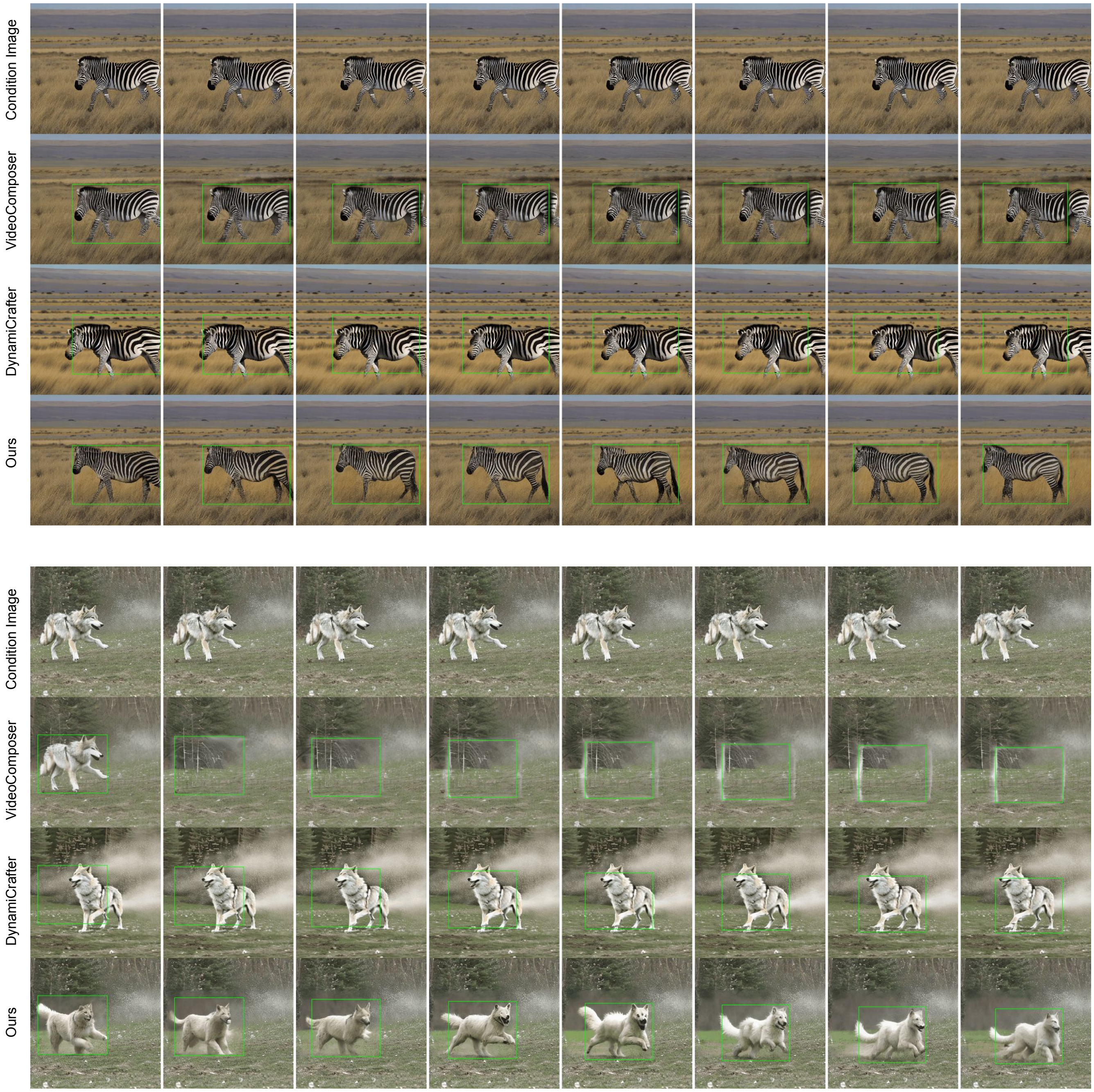


Figure 7: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.

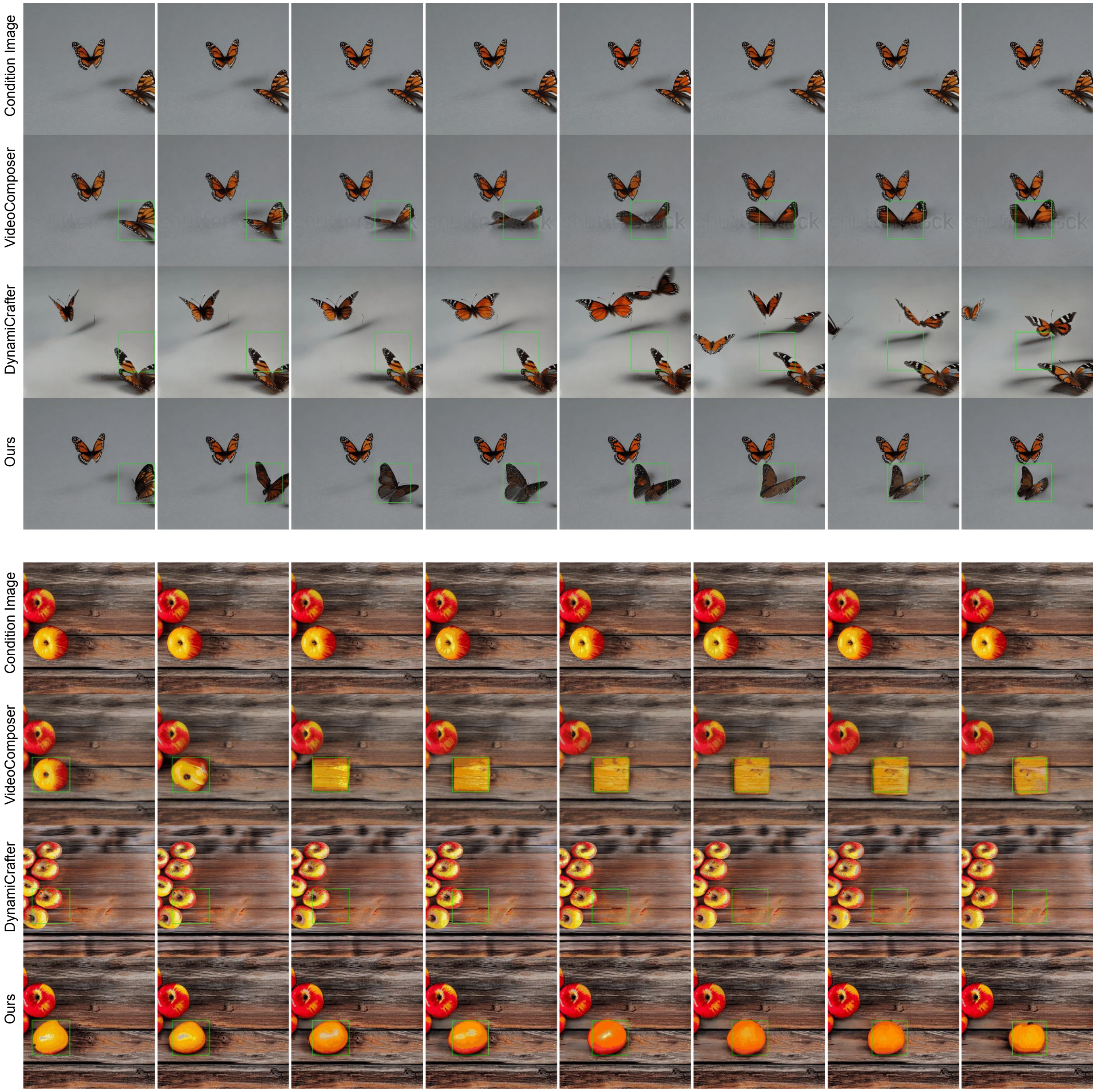


Figure 8: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.