Zero-Shot Controllable Image-to-Video Animation via Motion Decomposition

Shoubin Yu
shoubin@cs.unc.edu
UNC Chapel Hill

Jacob Zhiyuan Fang
zyfang@amazon.com
Amazon

Jian Zheng
nzhengji@amazon.com
Amazon

Gunnar Sigurdsson
gunnar@amazon.com
Amazon

Vicente Ordonez
vicenteor@rice.edu
Rice University

Robinson Piramuthu
robinpir@amazon.com
Amazon

Mohit Bansal
mbansal@cs.unc.edu
UNC Chapel Hill

ABSTRACT

In this paper, we introduce a new challenging task called Zero-Shot Controllable Image-to-Video Animation, where the goal is to animate an image based on motion trajectories defined by the user, without fine-tuning the base model. Primary challenges include maintaining consistency of background, consistency of object in motion, faithfulness to the user-defined trajectory, and quality of motion animation. We also introduce a novel approach for this task, leveraging diffusion models called Img2VidAnim-Zero (IVA^0). IVA^0 tackles our controllable Image-to-Video (I2V) task by decomposing it into two subtasks: 'out-of-place' and 'in-place' motion animation. Due to this decomposition, IVA^0 can leverage existing work on layout-conditioned image generation for out-of-place motion generation, and existing text-conditioned video generation methods for in-place motion animation, thus facilitating zero-shot generation. Our model also addresses key challenges for controllable animation, such as Layout Conditioning via Spatio-Temporal Masking to incorporate user guidance and Motion Afterimage Suppression (MAS) scheme to reduce object ghosting during out-of-place animation. Finally, we design a novel controllable I2V benchmark featuring diverse local- and global-level metrics. Results show IVA^0 as a new state-of-the-art, establishing a new standard for the zero-shot controllable I2V task. Our method highlights the simplicity and effectiveness of task decomposition and modularization for this novel task for future studies. Our code and visualizations are available at https://img2vidanim-0.github.io/

CCS CONCEPTS

• **Computing methodologies** → *Computer vision tasks.*

KEYWORDS

Image-to-Video Animation, Controllable Video Generation

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-0686-8/24/10.
https://doi.org/10.1145/3664647.3681394

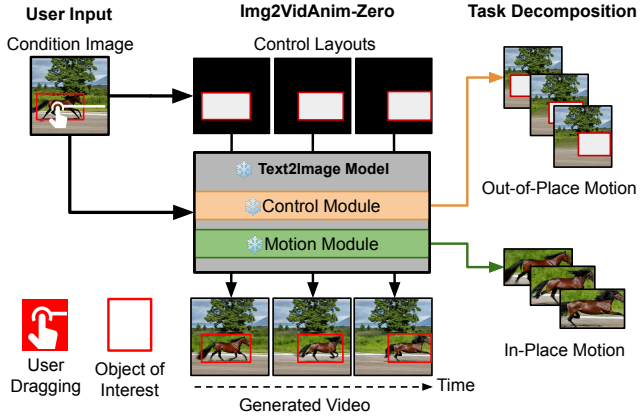

Figure 1: Img2VidAnim-Zero for zero-shot image-to-video animation based on motion trajectories from the user. We decompose controllable Image-to-Video generation as two subtasks: (1) out-of-place motion generation, and (2) in-place motion animation which can be solved by leveraging existing modules pre-trained on other task-specific data.

ACM Reference Format:
Shoubin Yu, Jacob Zhiyuan Fang, Jian Zheng, Gunnar Sigurdsson, Vicente Ordonez, Robinson Piramuthu, and Mohit Bansal. 2024. Zero-Shot Controllable Image-to-Video Animation via Motion Decomposition. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3664647.3681394

1 INTRODUCTION

The rising demand for controllable video generation underscores the desire of users to create videos for an increasing list of applications, such as personalized advertisement, educational content generation, visualization of imagination through user-generated content in social media, and entertainment content such as a short movie, where precise control of motion may be desired. While recent developments have made commendable strides in video generation from text prompts [15, 17, 19, 58, 64, 91], most works

do not allow users to control the finer-grained details easily and interactively (*e.g.*by drawing trajectories or defining layouts).

Current developments in controllable video generation focus on the Image-to-Video generation task (I2V) [4, 8, 22, 36, 66, 74, 78, 89]. I2V starts from a given condition image, eliminating the ambiguity often encountered with Text-to-Video (T2V) generation, enabling more diverse video animation based on additional conditions (*e.g.*text [8, 74], trajectory [4, 78], or reference video [89]). As a result, I2V blends precision, versatility, and a more user-friendly setup, positioning itself as a promising direction for controllable video generation. Recent I2V methods have utilized models trained on massive data with pre-extracted motion features [36, 66, 74, 78, 89]. However, ensuing challenges arise on: i) Computational resources: Even with efficient schemes like parameter-efficient fine-tuning [21], training models to understand new control conditions (e.g., motion vectors, trajectories) remain resource intensive. ii) Data collection: Acquiring data with meticulously annotated conditions can be expensive. Given these escalating costs, a pressing question is: *Can we devise a more cost-effective but still controllable I2V model?*

In this paper, we present Img2VidAnim-Zero (**IVA^0**) for Zero-shot Controllable Image-to-Video Animation without any I2V training data. The input set consists of the condition image and motion trajectories for objects of interest, represented by sequences of bounding box layouts. Our approach, as illustrated in Fig. 1 is to simplify the controllable I2V task by decomposing it into two atomic tasks. (1) **Out-of-place Motion Generation** focuses on determining the coarse layout of objects' obvious displacement throughout the video frames and (2) **In-place Motion Animation** ensures consistency while facilitating plausible, smooth motion (pixel-level changes) for user-dragged in-box objects across the frames. We have identified that each atomic task can be tackled with pre-existing modules from the Latent Diffusion [51] family, leading to our goal of *Zero-shot* Controllable I2V Animation.

As shown in Fig. 1, our IVA^0 is based on 3 core components: (1) a pre-trained text-to-image model [51], (2) the Control Module (CM), and (3) Motion Module (MM). *Out-of-place motion* generation is formulated as a layout-to-image generation task, achieved by inserting Gated Self-Attention Layers [35] as a layout control module, leveraging bounding box layouts for precise object placements. We refer to this as the Control Module (CM). *In-place motion* animation is achieved by adopting Temporal Attention Layers [17], from the text-to-video generation task. We refer to this as the Motion Module (MM). It maintains the consistency of the selected objects by applying self-attention across frames, leading to a realistic and smooth transition of objects from one frame to the next. Notably, CM and MM are pre-trained on corresponding task-aligned datasets without any I2V-specific training. We further propose an efficient *Motion Afterimage Suppression* (MAS) scheme that generates frames via alternating different inpainting operations to reduce *afterimage*[1] hallucination objects that could be left trailing behind the motion trajectory, while maintaining a reasonable background.

Lastly, the proposed novel task requires a corresponding benchmark dataset for evaluation. While prior work [22] evaluates controllable I2V on synthetic datasets [14, 31] with limited quantitative metrics, we construct a new test bed with diverse objects, annotated control layouts, and more concrete metrics. We assess controllable I2V across local object aspects (including control accuracy, appearance and motion consistency, and object residual), global scene aspects (including scene consistency, and video quality), and human evaluation. We compare our IVA^0 with other strong I2V models [8, 66] that are end-to-end trained with massive data on the I2V task. This comprehensive evaluation reveals our zero-shot IVA^0 to be superior across 9 metrics, setting a new state-of-the-art on the proposed I2V benchmark. In summary, our main contributions are:

- Novel task of Zero-shot I2V Animation based on user-defined layout trajectories, along with novel approach Img2VidAnim-Zero (IVA^0)
- Novel controllable I2V benchmark, enriched with diverse visual content and annotated control layouts (will be released), evaluated models across diverse dimensions.
- The Img2VidAnim-Zero achieves competing results on the zero-shot I2V task. Our quantitative and qualitative results highlight the effectiveness and potential of our modular task decomposition idea for future controllable I2V studies.

2 RELATED WORKS

Video Generation. As a generative task with promising prospects, video generation has been a popular research topic. Early efforts [52, 61, 63, 82] focus on the unconditional generation that is based on the vector initialized from a pre-defined probability space (*e.g.*Gaussian distribution). Recent works introduce various generation conditions and can be roughly categorized into: i) Text-to-Video generation (**T2V**) [1, 6, 15, 17, 19, 24, 29, 37, 39, 58, 67, 91, 92]: where descriptive text is used as input to guide the generation process, ii) Video-to-Video generation (**V2V**) [9, 23, 42, 48, 70, 71, 77, 79]: wherein a reference video informs the structure of the generated video, and iii) Image-to-Video generation (**I2V**) [4, 8, 22, 28, 36, 56, 66, 74, 78, 84, 88, 93]: which uses a single or a series of images as the basis to produce a continuous frame sequence. We focus on the I2V formulation, which provides a clear visual starting point compared with T2V and gives more flexibility compared with V2V. Our work aims to inject layout-based controllability into the I2V.

Controllability in I2V Generation. Controllable video generation has become increasingly popular. Many efforts centered around encoding images and motion trajectories, mainly for human movement [2, 4, 5, 16, 72], editing a reference video through fine-grained control (e.g. dragging, depth/edge/pose maps) [13, 38, 43, 55, 60]. Recent advancements include [78], which allows fine-grained object motion through user-defined trajectories, leveraging extensive video data, extra motion feature extraction, and multi-scale fusion modules. [66] is adept at synthesizing videos based on various combinations of appearance and motion patterns, given its training with varied spatio-temporal conditions. [36] introduces a neural stochastic motion texture for still images, ideal for objects with limited motion. Latest developments like [25, 37, 39] incorporate LLM layout planning into generation (and the former also introduces consistency in long-video generation) but these works focus on T2V generation. Very recent studies [10, 12, 26, 50, 65, 68, 69, 71, 76] make great progress in fine-tuning the model to be aware of diverse control conditions (e.g. detailed textual prompts, trajectories, boxes, and reference video) to animate an image. Furthermore, some other

[1] https://en.wikipedia.org/wiki/Afterimage

recent works adopt the efficient zero-shot setting, but they are suffering from controllability [80], focusing on T2V [6] or with extra LLM planning [59], or conducting extra DDIM operation [7]. Our method is not text-based or fine-tuned for the task, but an image-based zero-shot generation without any I2V-specific training or DDIM inversion. In addition, our method can also combine with these LLM planners to generate layout trajectory/sequence based on text for a more diverse control condition input.

3 METHOD

In this section, we introduce our IVA^0 model in detail. We first discuss the formulation of inpainting based on latent diffusion [51], which serves as the foundation of our model (Sec. 3.1). We then present how we build up IVA^0 by decomposing controllable I2V into sub-tasks, which can be addressed with out-of-place and in-place motion modules in the Latent Diffusion family (Sec. 3.2).[2] Finally, we elaborate on our Motion Afterimage Suppression (MAS) schema to eliminate object afterimage hallucination (Sec. 3.3).

3.1 Preliminary: Latent Diffusion for Inpainting

Our objective is to animate a static object, facilitating its transition from an initial to the subsequent position based on any user-defined layout trajectory. We frame it as an inpainting task for controlling such object movement, which involves: (1) replacing the original object location with the background, while (2) inpainting the object based on layout. To accomplish this, our IVA^0 is constructed on the inpainting version of Latent Diffusion [51], a publicly available pre-trained text-to-image model. The Latent Diffusion model comprises three key components: (1) Autoencoder that maps the image from pixel space to latent embedding, based on which the Diffusion module operates; then projects the embedding after denoising steps back to pixel space; (2) Text encoder that encodes a prompt into embedding for Text-to-Image conditioning; (3) U-Net for noise diffusion, which iteratively conducts denoising in the latent space, guided by timestamps and prompt embedding.

The inpainting task leverages the Latent Diffusion model to modify a masked image region based on the given textual conditions. This mask is represented as a 1-channel binary mask as additional input together with the condition image. The latter delivers essential context for the un-inpainted sections and is derived by processing a condition image x_{con} through the encoder. To adapt to these extra inpainting conditions, the diffusion U-Net incorporates five extra channels in its initial convolution layer. Given a condition image x_{con}, a text prompt p, and a binary mask m, the inpainting model generates an image. In the following sections, as depicted in Fig. 2, we delineate the integration of control conditions and how we handle basic atomic tasks with this text-to-image inpainting model to achieve controllable I2V.

3.2 Zero-Shot Layout-Conditioned I2V

In IVA^0, we introduce a controllable Image-to-Video generation model that leverages user-provided spatio-temporal object layouts. Given an initial frame x_1 as condition image x_{con}, users can animate specific objects by providing a trajectory of the object. This trajectory, in our method, is represented as a sequence of bounding boxes: $(b_1 \cdots b_t)$, where t refers to the number of frames. Each box b_i is a 4-dimensional vector indicating the top-left and bottom-right coordinates of the box. For simplicity, we focus on animating only a single object at a time. But, IVA^0 is versatile and can be extended to multiple objects simultaneously when provided with corresponding layouts. The detailed model pipeline is elaborated as follows:

Layout Condition via Spatio-temporal Masking: A future frame x_i is generated based on the initial frame x_1 (x_{con}) and layout boxes via the T2I model. When transitioning the object from its position b_1 to b_i, we expect the object to move to the desired location with smoothly interpolated motion and consistent appearance for both the foreground object and background context. This necessitates inpainting x_i in two regions: (1) eliminate the object from the original region at b_1 as part of the background, and (2) add the object to the new region b_i. So we create an inpainting mask for each frame by simultaneously masking out both starting region b_1 and target region b_i. As illustrated in the left of Fig. 2, from the spatio-temporal layout sequence $(b_1 \dots b_t)$, we construct the spatio-temporal masking sequence $\mathbf{M} = (m_1 \dots m_t)$, with each $m_i = b_1 \cup b_i$.

Out-of-Place Motion Generation: An important task of our model is to generate out-of-place motion, given the spatio-temporal masks $\mathbf{M}$. This can be formulated as a layout-conditioned generation similar to [11, 35, 87, 90], which requires generating an image following a layout condition. Thus, we adopt the design of gated self-attention proposed in GLIGEN [35], a layout-to-image generation model. GLIGEN encodes object box coordinates into special *grounding tokens* and fuses grounding information with visual tokens via extra gated self-attention that is added before each cross-attention layer in the text-to-image model. Specifically, as shown in the middle of Fig. 2, we insert Gated Self-Attention layers inherited from [35] with copied weights, as our Control Module. The control module then utilizes grounding tokens that encapsulate both the appearance of the object and its box coordinates, enabling precise placement of the object in the desired location. We streamline the process by using the same CLIP model [49] as an image encoder to extract regional image features of the cropped object. The box coordinates b_i are projected into continuous embedding with Fourier transform function as in [44], controlling the spatial location. Thus, for the frame at time i, layout tokens h_i are derived by integrating these conditions via a linear projection layer. These tokens then interact with the visual tokens of the same frame using gated self-attention, ensuring accurate and contextually relevant out-of-place motion generation, such that[3]:

$$h_i = MLP(CLIP_{\text{img}}(crop(x_1, b_1)), Fourier(b_i)) \tag{1}$$

$$v_i = SelfAtt(concat(h_i, v_i)) \tag{2}$$

In-Place Motion Animation: Based on our observation, relying solely on the out-of-place inpainting strategy only produces a rudimentary "copy-paste" animation for objects (see Fig. 6), causing noticeable inconsistencies in their motions across frames. In order to pursue a smoother and authentic object-moving motion and ensure sustained visual coherence, we adopt an in-place motion animation module. Previous works [15, 29, 64, 66] show different

[2] In our implementation, we adopt Stable Diffusion which is an improved variance of the Latent Diffusion Model: https://github.com/Stability-AI/StableDiffusion.

[3] Here the concatenation is for one attention block. Details in [35].

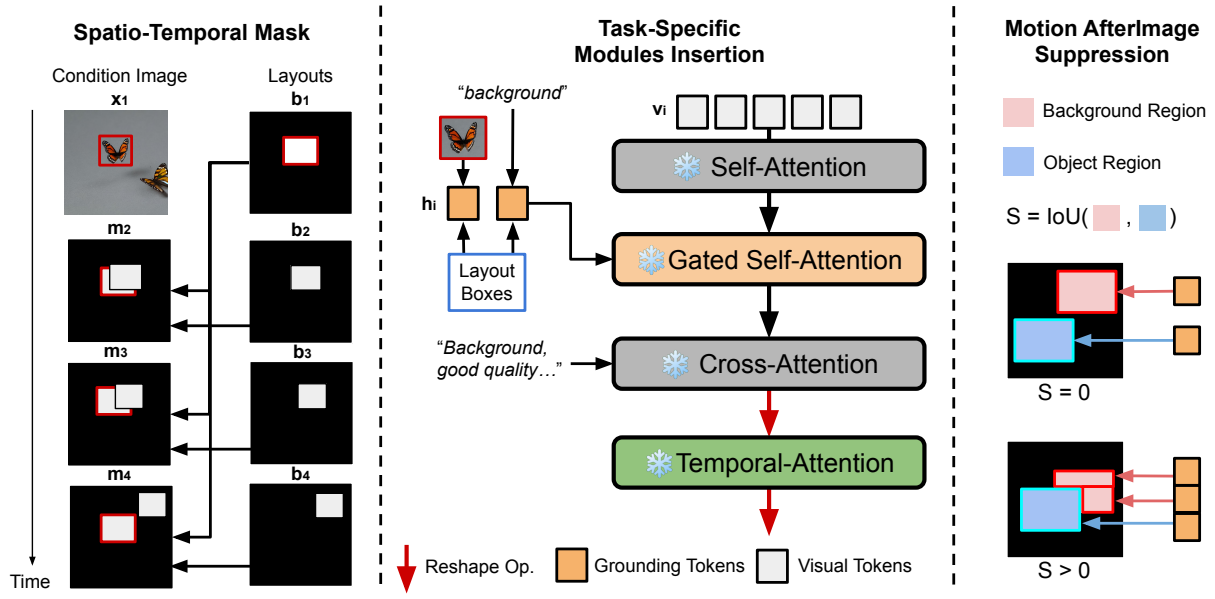

Figure 2: Left: Our spatio-temporal inpainting masks cover the starting position of objects and their target locations in each frame. Middle: To handle various atomic tasks in controllable I2V, we integrate different task-specific modules. We apply gated self-attention layers as the control module for generating out-of-place motion, while using temporal-attention layers as the motion module for in-place motion animation. Right: We introduce Motion Afterimage Suppression (MAS), which uses object size and IoU to decide whether to inpaint the background with additional grounding tokens. This approach aims for enhanced inpainting quality with reduced afterimage hallucination.

inter-frame attention mechanism that helps this goal, but unanimously require large-scale pre-training from video data. We resort to a pre-trained motion engine [15], as illustrated in the middle of Fig. 2, and incorporate its Temporal Attention layers [15] with weights copied from the original Text-to-Video generation task, but for our controllable I2V task. This motion module enables better temporal consistency for both object appearance and motion via self-attention across frames. Specifically, given the sequential frame visual features $\mathbf{V} = (v_1 \dots v_t)$, where $\mathbf{V} \in R^{(t,h*w,c)}$, we reshape the feature axes and apply self-attention to the temporal dimension, where w, h, and c refer to width, height, and feature channel.

$$\mathbf{V} = Reshape(SelfAtt(Reshape(\mathbf{V}))) \tag{3}$$

Both the control and motion modules are pre-trained on different task-specific data. They integrate capabilities from their original Layout-Conditioned Image Generation and Text-to-Video tasks. We incorporate these established foundations for our generation to avoid further re-training. Thus, our IVA^0 can controllably animate objects in an image without any Controllable I2V-specific fine-tuning. Sec. 4.2 contains more training details of these modules.

3.3 Motion Afterimage Suppression

As our IVA^0 model is built upon the Text-to-Image inpainting model, we initially prompt the model with a fixed background-filling text prompt (shown in Fig. 2 middle) for background generation. However, as illustrated in Fig. 3 bottom-middle, we observed this approach sometimes results in an afterimage, a ghost-like residual hallucination after the object has moved (bottom row of Fig. 3). We

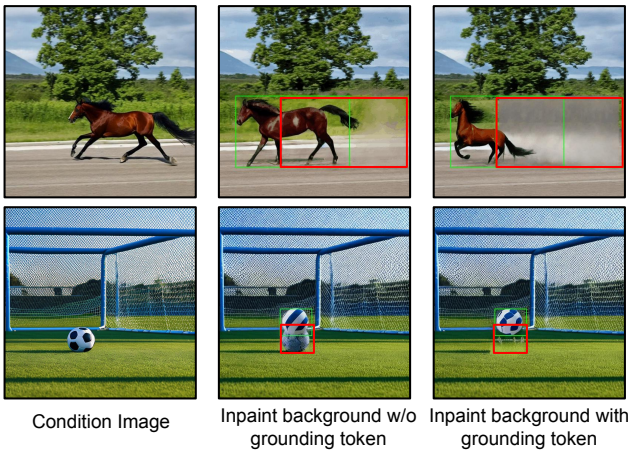

Figure 3: Comparison of background inpainting with or without grounding tokens. The token-based method can eliminate the afterimage object hallucination issue (middle-bottom red box) but is weaker in handling larger regions (right-top red box) compared with the token-free one.

further find out that this issue is linked to the motion module in our model. As evidenced in Fig. 6 (rows 2 and 3), these hallucinations occur when the motion module is used in the model. This is because most current motion modules [8, 17, 74] only generate limited-range, in-place motion. For this reason, when an object

moves significantly from its original b_1, the temporal attention fails to maintain appearance consistency. Inversely, the temporal attention often wrongly produces an afterimage at b_1 in x_t.

To suppress such a motion afterimage, we experimented with using extra grounding tokens for the background generation. As illustrated in Fig. 2 middle, we set grounding tokens that encode both the bounding box and the word 'background' (via CLIP text embedding). In this case, we force the control module to generate a background at b_1 in frame x_t. This method successfully avoids afterimage hallucinations for small objects but struggles with large areas (see Fig. 3 top-right). This limitation likely stems from the control module's training on background reconstruction, which cannot handle large-region in-painting with a single token. To overcome this, we integrated two background generation approaches based on object size and Intersection over Union (IoU). Objects are first categorized by size (small, medium, large) with pre-defined area thresholds. Small objects use extra grounding tokens for background in-painting, whereas large objects do not. For medium-sized objects, as shown in Fig. 2 (right), we first calculate the IoU S between b_1 and b_t, if $S > 0$, indicating overlap, the non-overlapping background areas are first divided into grids iteratively. The model then in-paints each grid with background class tokens. Experiments show that our proposed MAS resolves the afterimage object hallucination while maintaining high-quality background generation.

4 EXPERIMENTS

Our experimental setup is detailed in this section, including proposed metrics & benchmark for quantitative evaluation (Sec. 4.1), implementation details (Sec. 4.2), results analysis (Sec. 4.3), and limitation discussion (Sec. 4.4). Details of benchmark data collection, human evaluation, and baseline implementation are in Appendix.

4.1 Evaluation

To comprehensively assess the performance of our model under controlled experimental conditions, we evaluate its effectiveness across various metrics. This evaluation encompasses both the local level (concentrating on objects achieving smooth motion, maintaining a consistent appearance, and following layout conditions), and the global level (focusing on generated video scenes matching the given condition image).

- **vIoU@R:** This evaluates the correctness of layout conditioning. We adopt this metric from action detection [75, 86], which calculates the spatio-temporal overlap between ground truth and GroundingDINO [41] detected boxes: i.e., if vIoU exceeds threshold R, then we consider the prediction to be a match, namely vIoU@R. In our experiment, we report results with R = 0.3, 0.5.
- **Smoothness:** We propose the "Smoothness" metric, which evaluates the smoothness of object transformations across video frames. We compare object similarity in consecutive frames using GroundingDINO for object detection and CLIP ViT-B/32 for image embeddings. Smoothness is calculated by averaging the similarity of these embeddings across all frames, with a higher score indicating smoother changes in appearance in the video.
- **Hallucination:** We propose the "Hallucination" metric, capturing the wrong afterimage generations. We first detect the target object class in each frame with GroundingDiNO. Then, we calculate the difference in object count between the generated and the condition image and sum normalized results over multiple frames for a video-level metric. This metric reflects the unexpected generation (e.g., extra object) or removal of an object.
- **SSIM & LPIPS:** These measure the structural similarity of generated frames with the condition image. As we are only interested in animating the object in the image while maintaining other regions, the higher structural similarity between the condition image and the generated frames means that the model can keep the background scene or non-interested regions unchanged. We adopt both non-parametric SSIM [20] and parametric LPIPS [83] to represent structural similarity.
- **FID & FVD:** These standard reference-based metrics ([18], [62]) quantify the visual appeal by comparing the sets of ground truth and generated videos. To compute FVD, we repeat and stack initial frames as pseudo-video for distribution gap computation.
- **Human Evaluation:** Automated metrics are not perfect. Hence, we include human evaluation as well. Annotators are asked to conduct a majority voting for the best-quality video considering the appearance consistency of the object/background, motion faithfulness, and motion quality.

Controllable I2V Benchmark: Since our controllable I2V task focuses on the animation of objects, we evaluate the model on a testbed that contains ground-truth videos without camera motion, involves diverse objects, permits reasonable motion ranges, and includes controlled layouts. For this:

- We collect 100 images as initial frames, a mix of generated ones using Stable Diffusion [51], and real images from a public dataset [47]. These images feature diverse objects for animation.
- For each image, we manually annotate the start and end boxes for an object and interpolate them with intermediate boxes using a non-linear function as a trajectory.
- We also annotate each sample with a textual prompt, describing the desired motion of the object. In total, our testing set comprises 200 object control layouts & captions, paired with 100 images.

Baseline Models: We compare our method against two competing image-to-video generation baselines: 1) *VideoComposer* [66], a compositional video synthesis model that offers motion controllability conditioned on motion vector. It is pre-trained on *WebVid10M* [3] and *LAION-400M* [53]. It is based on the Video Latent Diffusion Model (VLDM) [64] that incorporates both 3D convolution and temporal attention. In the evaluation, we adapt our layouts into motion vector format for compatibility. 2) *VideoCrafter* [8], a recent emerging image animation model that has been pre-trained on *WebVid10M* dataset [3]. It is also built on the VLDM. Since *VideoCrafter* is unable to be conditioned by trajectory, we circumvent this by providing an extremely detailed prompt as the condition.

4.2 Implementation Details

Model Implementation: We choose Stable Diffusion-v1.4 as our base model, which is pre-trained on *LAION-400M* [53]. Our control module is derived from GLIGEN [35], parameter-efficient [73] pre-trained on various grounding datasets, including *COCO2014D, COCO2014CD, and COCO2014G.* Our motion module, derived from AnimateDiff [15], is pre-trained on *WebVid10M* [3]. The guidance

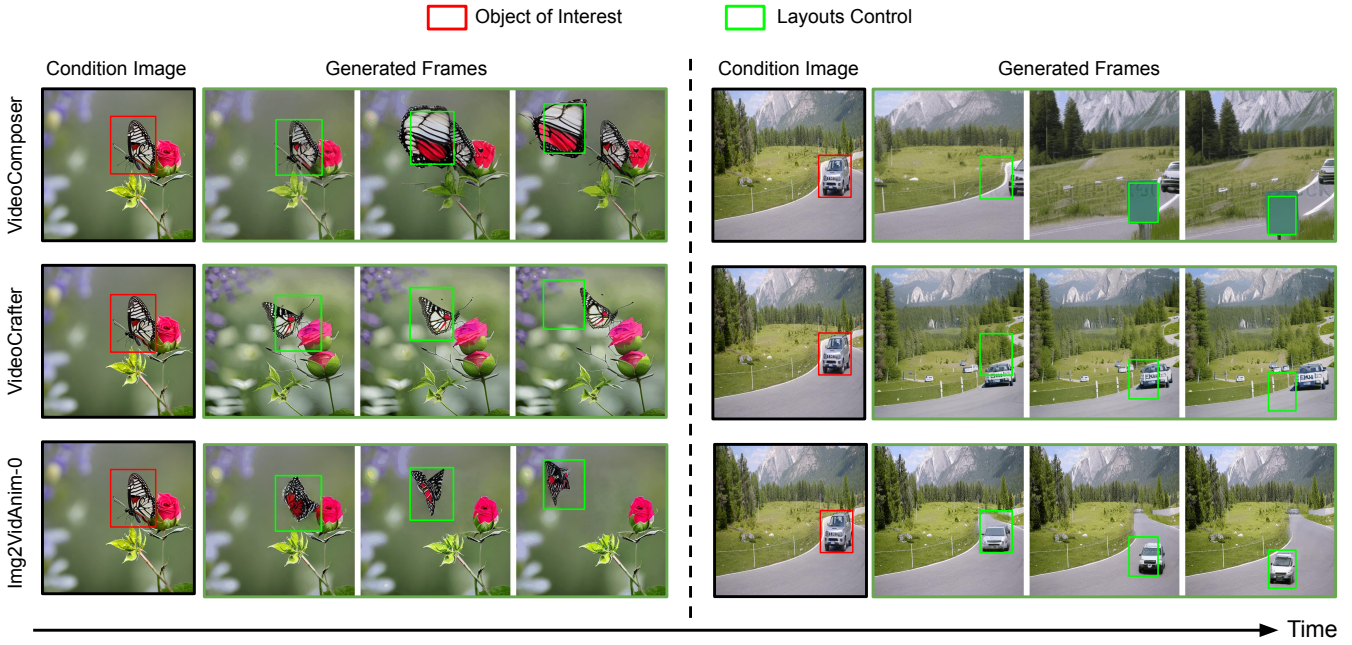

Figure 4: Qualitative comparison of our method and baseline approaches. The sample on the left is based on initialization with a generated image, while the sample on the right was initialized with a real image. Best viewed in color and zoomed in for more details.

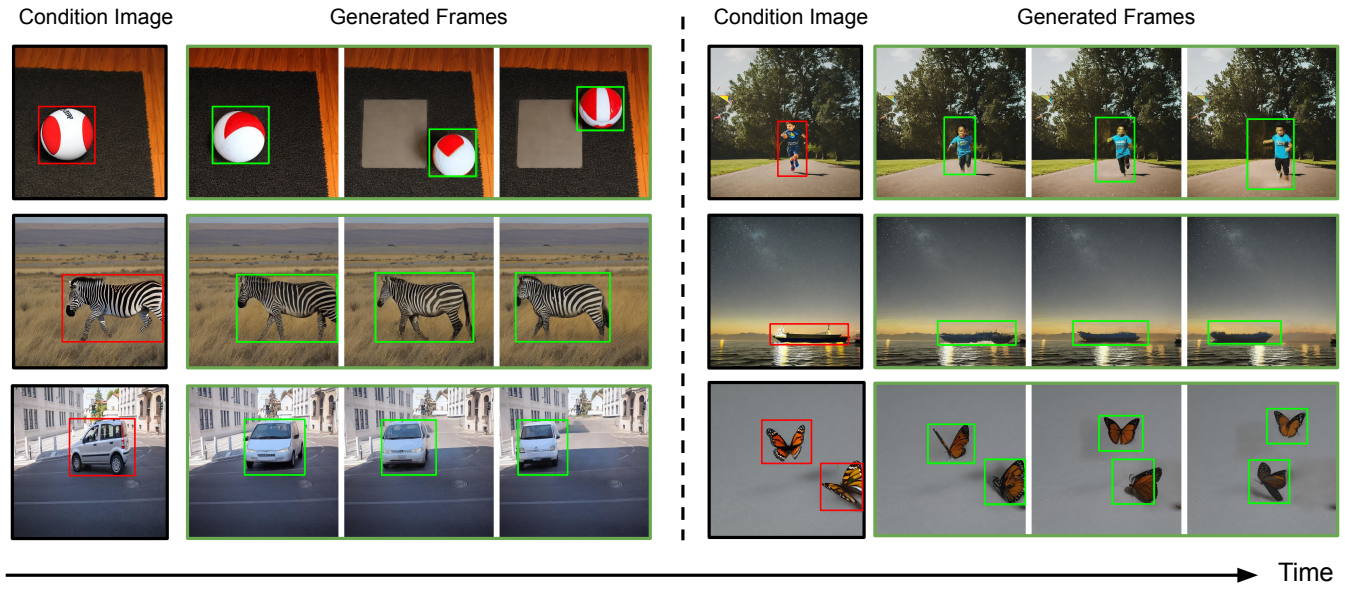

Figure 5: Sample results of the proposed IVA^0. The green boxes in the generated frames represent input control layouts. Best viewed in color and zoomed in for the best visualization. We provide more qualitative results in the Appendix.

scale was set to 7.5 and 25 steps were used for denoising with DDIM noise scheduler. 16-frame videos were generated for both our method and baseline methods. The resolution was set at 512×512 for generated videos. 5 random seeds were used for each layout trajectory; We categorize objects based on their area relative to the image: small objects occupy no more than 1/16, medium ones range between 1/16 and 1/5, and large ones exceed 1/5 of the image area. We use 5 random seeds (1, 2, 3, 42, 126) to get five unique generations for each image-box pair. Our noise scheduler adopts 0.00075 beta start, 0.012 beta end, and "scaled_linear" beta schedule.

Table 1: Quantitative comparison with evaluated baselines on the proposed controllable I2V benchmark. Smooth.: object smoothness across frames. Hallu.: object hallucination in generated videos.

Methods	0-Shot	Local Object				Global Scene				HumanEval
		vIoU@0.3 ↑	vIoU@0.5 ↑	Smooth. ↑	Hallu. ↓	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	Win Rate↑
VideoComposer [66]	N	50.0	25.3	85.8	42.4	36.4	45.2	151.4	1790	31%
VideoCrafter [8]	N	21.4	5.4	89.2	42.5	25.9	56.0	**132.1**	1460	27%
IVA^0 (ours)	Y	**88.7**	**78.1**	**90.2**	**13.6**	**61.9**	**28.0**	132.7	**1352**	**42%**

The negative prompt used for the generation was: "worst quality, deformed, extra object, extra human, distorted, disfigured, bad anatomy, disconnected limbs, wrong body proportions, low quality, illustration, oversaturated, cartoons, blurry, cropped, text."

Baseline Implementation: For VideoCrafter [8], which does not support conditioning with layout trajectories, we prompt the model with the detailed motion-related text (see Appendix). For VideoComposer [66], which can use hand-crafted motion vectors as extra conditions, we first generate motion vectors from layout sequences and then use the vectors to generate video.

Data Collection: We collect both 20 real-world images from public tracking/segmentation datasets [47] and 80 generated images by StableDiffusion-2.0 [51]. We manually write text prompts for the StableDiffusion model to generate images. We collect images containing single/multiple objects and with an empty area that allows for obvious out-of-place object motion.

Data Annotation: To reduce the labeling burden, we only manually annotate key layout boxes for each image, and interpolate those key boxes to imitate user dragging and obtain layout sequence during inference. We also write detailed motion-related captions for each image-layout pair to prompt baseline models [8] that cannot take layout input. More details are in Appendix.

4.3 Results

Qualitative Results: Fig. 4 shows two examples of generation with our method compared with baselines. In practice, we find it challenging to leverage *VideoCrafter* to animate the object strictly following user instructions by purely prompting; e.g., the Lepidoptera and car can barely follow the box condition. Though *VideoComposer* shows good layout conditioning, it suffers from the hallucination of new objects or sometimes the removal of the target object; e.g., an extra Lepidoptera was not successfully eliminated in the original region. We find our IVA^0 follows the layout condition better than the baselines and also generates smooth motion with a more consistent object appearance. We also provide more samples generated by our IVA^0 in Fig. 5 with various animation conditions, including single/multiple objects, small/large objects, real/generated images, and simple/complex motion trajectories. We observe consistent conclusions across all combinations. However, we do observe all methods performing less satisfactorily for maintaining consistent object appearance across frames; e.g., in Fig. 5, appearance of children and vehicles is altered in the generated frames. We conjecture for these reasons: 1. It still remains challenging for the T2I model to produce customized & consistent objects without any weight optimization; 2. CLIP embedding does not capture information about the object comprehensively.

Quantitative Results Analysis: Tab. 1 contains quantitative results. Our IVA^0 shows leading results on local-object metrics, i.e., vIoU, Smoothness, and Hallucination. Specifically, IVA^0 largely surpasses *VideoComposer* and *VideoCrafter*: 88.7 vs. 50.0 and 21.4 when R=0.3. This verifies that our IVA^0 possesses the capability of precise layout control. In addition, we notice that IVA^0 has a higher Smoothness score: 90.2 vs. 89.2 by *VideoCrafter*, indicating that our IVA^0 produces objects with smoother changes across frames. Noticeably, the obvious lead in the Hallucination metric score (13.6 vs. 42.5) also consolidates this conclusion. When evaluated at the scene level, we observe very aligned results. IVA^0 leads the baselines on SSIM, LPIPS, and FVD, while showing a close gap with *VideoCrafter* on FID: 132.7 vs. 132.1. We note that both FID and FVD compute the distribution gap. Since our evaluation is just based on a single initial frame, the slight gap here just indicates that the distribution of the produced video frames is not evidently deviating from the initial frame. Furthermore, our human evaluation study shows the superiority of the animation produced by our method: 42% win rate versus 31% and 27% for both baselines. Details on human studies can be found in the Section 4.2.

Ablation Study. We now assess the effectiveness of each module and the Motion Afterimage Suppression strategy. As illustrated in Fig. 6, we observe that CM largely improves the control ability of IVA^0 on vIoU@0.3: 26.7 vs. 88.7. Adopting a motion module explicitly improves motion smoothness, which is reflected in the Smoothness and FVD metrics. We observe that the motion module has negative impacts on FID, LPIPS metrics, which aligns with our expectations as it is interpolating frames with diverse motion, inevitably diverging it from the initial frame used as ground truth on these metrics. The MAS module largely improves the baseline performances on a series of local object metrics: 88.7 vs. 87.3 on vIoU@0.3 and 13.6 vs. 21.2 on the Hallucination metric. It should be noted that while integrating the motion module improves temporal video metrics like Smoothness and FVD, it also diminishes the grounding capability (measured by VIoU) facilitated by the control module. It is a trade-off for such a zero-shot model without extra alignment for injected modules. The full model achieves a synergistic balance between these capabilities across modules, leading to better overall quality. Fig. 6 further showcases more examples of generated images with each module.

4.4 Limitations & Discussion

Despite of the generated video, our zero-shot model IVA^0 still faces several challenges, as shown in Fig. 7, as following:

(1) Inconsistent & Missing Object: IVA^0 struggles to maintain appearance consistency of object appearance w.r.t. target object

Table 2: Ablation study on our method. CM: Control Module. MM: Motion Module. MAS: Motion Afterimage Suppression.

MM	CM	MAS	Local Object				Global Scene			
			vIoU@0.3 ↑	vIoU@0.5 ↑	Smooth. ↑	Hallu. ↓	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
-	✓	✓	**90.2**	**85.7**	88.4	14.4	**63.2**	**25.9**	**118.4**	1440
✓	-	✓	26.7	21.6	70.1	43.5	58.4	32.8	182.0	1388
✓	✓	-	87.3	77.5	89.5	21.2	62.4	27.4	133.6	1369
✓	✓	✓	88.7	78.1	**90.2**	**13.6**	61.9	28.0	132.7	**1352**

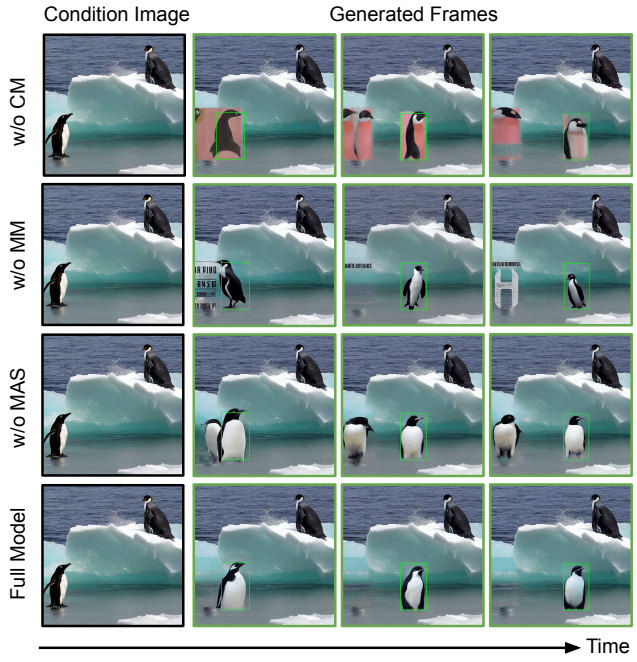

Figure 6: Qualitative images are produced by our IVA^0 with and without each module. The last row shows the images produced by the combination of all modules.

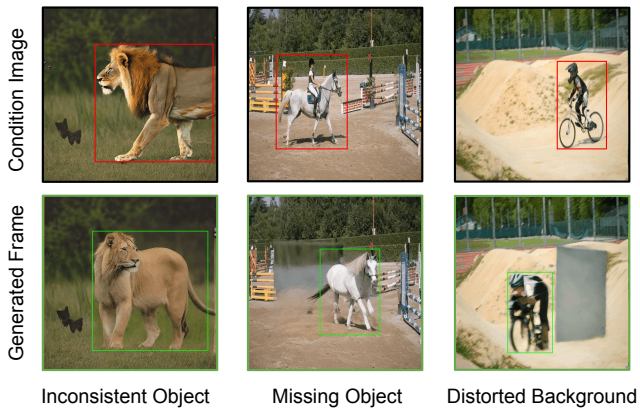

Figure 7: Examples of failure cases in our model include producing objects with an inconsistent appearance, missing objects, or distortions in the background.

in the initial frame, and it tends to overlook secondary objects when multiple objects are present in a single box (*e.g.*woman on the horse). This discrepancy arises possibly because CLIP image embedding adeptly captures mainly high-level features but neglects the finer low-level features (i.e., texture, color, and 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 [45]. 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 details. **(2) Distorted background**: it inaccurately generates backgrounds (*e.g.*gray patch behind the man on bike), especially for real images with complex scenes. We attribute those to a lack of extra training/alignment with the pre-trained motion and control modules for the base T2I model. It weakens the base model's inpainting and generation ability. This mismatch is further exacerbated when applied to real-world images due to a potential domain gap in pre-trained models [34, 45]. Besides, our model is currently limited to animating foreground objects, and cannot modify the background in case users want to animate the background too or incorporate camera motion.

However, we note that the spatio-temporal consistency across scenes and objects still remains an open challenge for all existing T2V/I2V models, which rely on large-scale pre-training as a possible solution [46]. Our efforts focus on building an efficient zero-shot T2V model without any tuning. As part of future directions, we suspect (1) integration with a text-to-video backbone that applies temporal modules, e.g., 3D convolution, (2) fine-tuning with I2V data will further enhance consistency, (3) integration with more control conditions [38, 85], e.g., segmentation masks. We also attempt to mitigate those bad generations with recent popular one/few-shot tuning ideas in text-to-video generation work (e.g., [70, 71]) (see more details in Appendix) and give more insights. In all, addressing these challenges remains an open topic for future research.

5 CONCLUSION

In this paper, we introduced Img2VidAnim-Zero (IVA^0), a Zero-shot Image-to-Video (I2V) method without task-specific I2V training. By harnessing existing text-to-image Diffusion modules and integrating Gated and Temporal Attention layers, IVA^0 facilitates accurate and seamless video generation from the specified motion trajectory based on bounding boxes. Our novel I2V benchmark underscores IVA^0's leading performance, showcasing its potential for future video generation applications.

REFERENCES

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. 2023. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477* (2023).

[2] Pierfrancesco Ardino, Marco De Nadai, Bruno Lepri, Elisa Ricci, and Stéphane Lathuilière. 2021. Click to move: Controlling video generation with sparse motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14749–14758.

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1728–1738.

[4] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. 2021. ipoke: Poking a still image for controlled stochastic video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14707–14717.

[5] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Bjorn Ommer. 2021. Understanding object dynamics for interactive image-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5171–5181.

[6] Ryan Burgert, Kanchana Ranasinghe, Xiang Li, and Michael S Ryoo. 2022. Peekaboo: Text to image diffusion models are zero-shot segmentors. *arXiv preprint arXiv:2211.13224* (2022).

[7] Changgu Chen, Junwei Shu, Lianggangxu Chen, Gaoqi He, Changbo Wang, and Yang Li. 2024. Motion-Zero: Zero-Shot Moving Object Control Framework for Diffusion-Based Video Generation. *arXiv preprint arXiv:2401.10150* (2024).

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. 2023. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. arXiv:2310.19512 [cs.CV]

[9] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. 2023. Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models. *arXiv preprint arXiv:2305.13840* (2023).

[10] Xi Chen, Zhiheng Liu, Mengting Chen, Yutong Feng, Yu Liu, Yujun Shen, and Hengshuang Zhao. 2023. LivePhoto: Real Image Animation with Text-guided Motion Control. *arXiv preprint arXiv:2312.02928* (2023).

[11] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. 2023. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908* (2023).

[12] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. 2023. AnimateAnything: Fine-Grained Open Domain Image Animation with Motion Guidance. *arXiv e-prints* (2023), arXiv–2311.

[13] Yufan Deng, Ruida Wang, Yuhao Zhang, Yu-Wing Tai, and Chi-Keung Tang. 2023. Dragvideo: Interactive drag-style video editing. *arXiv preprint arXiv:2312.02216* (2023).

[14] Rohit Girdhar and Deva Ramanan. 2019. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. *arXiv preprint arXiv:1910.04744* (2019).

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

[16] Zekun Hao, Xun Huang, and Serge Belongie. 2018. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7854–7863.

[17] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940* (2023).

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, Vol. 30.

[19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

[20] Alain Hore and Djemel Ziou. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*. IEEE, 2366–2369.

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

[22] Yaosi Hu, Chong Luo, and Zhenzhong Chen. 2022. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18219–18228.

[23] Zhihao Hu and Dong Xu. 2023. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073* (2023).

[24] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. 2023. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *arXiv preprint arXiv:2309.14494* (2023).

[25] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. 2024. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems* 36 (2024).

[26] Hsin-Ping Huang, Yu-Chuan Su, Deqing Sun, Lu Jiang, Xuhui Jia, Yukun Zhu, and Ming-Hsuan Yang. 2023. Fine-grained controllable video generation via object appearance and context. *arXiv preprint arXiv:2312.02919* (2023).

[27] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1780–1790.

[28] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. 2023. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025* (2023).

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

[30] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision* 128, 7 (2020), 1956–1981.

[31] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989), 541–551.

[32] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. 2022. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems* 35 (2022), 9287–9301.

[33] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10965–10975.

[34] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. 2023. StyleDiffusion: Prompt-Embedding Inversion for Text-Based Editing. *arXiv preprint arXiv:2303.15649* (2023).

[35] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.

[36] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. 2023. Generative Image Dynamics. *arXiv preprint arXiv:2309.07906* (2023).

[37] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. 2023. LLM-GROUNDED VIDEO DIFFUSION MODELS. *arXiv preprint arXiv:2309.17444* (2023).

[38] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. 2024. Ctrl-Adapter: An Efficient and Versatile Framework for Adapting Diverse Controls to Any Diffusion Model. *arXiv preprint arXiv:2404.09967* (2024).

[39] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. 2023. VideoDirectorGPT: Consistent Multi-scene Video Generation via LLM-Guided Planning. *arXiv preprint arXiv:2309.15091* (2023).

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

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

[43] Yue Ma, Xiaodong Cun, Yingqing He, Chenyang Qi, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. 2023. MagicStick: Controllable Video Editing via Control Handle Transformations. *arXiv preprint arXiv:2312.03047* (2023).

[44] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[45] 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.

[46] openai. 2024. https://openai.com/sora.

[47] 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.

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

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[50] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. 2024. ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation. *arXiv preprint arXiv:2402.04324* (2024).

[51] 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.

[52] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. 2017. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*. 2830–2839.

[53] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).

[54] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8430–8439.

[55] Fengyuan Shi, Jiaxi Gu, Hang Xu, Songcen Xu, Wei Zhang, and Limin Wang. 2023. BIVDiff: A Training-Free Framework for General-Purpose Video Synthesis via Bridging Image and Video Diffusion Models. *arXiv preprint arXiv:2312.02813* (2023).

[56] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. 2024. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

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

[58] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).

[59] Sitong Su, Litao Guo, Lianli Gao, Hengtao Shen, and Jingkuan Song. 2023. MotionZero: Exploiting Motion Priors for Zero-shot Text-to-Video Generation. *arXiv preprint arXiv:2311.16635* (2023).

[60] Yao Teng, Enze Xie, Yue Wu, Haoyu Han, Zhenguo Li, and Xihui Liu. 2023. Drag-A-Video: Non-rigid Video Editing with Point-based Interaction. *arXiv preprint arXiv:2312.02936* (2023).

[61] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1526–1535.

[62] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new Metric for Video Generation. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop*.

[63] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. *Advances in neural information processing systems* 29 (2016).

[64] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. ModelScope Text-to-Video Technical Report. *arXiv preprint arXiv:2308.06571* (2023).

[65] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. 2024. Boximator: Generating Rich and Controllable Motions for Video Synthesis. *arXiv preprint arXiv:2402.01566* (2024).

[66] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. VideoComposer: Compositional Video Synthesis with Motion Controllability. *arXiv preprint arXiv:2306.02018* (2023).

[67] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. 2023. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103* (2023).

[68] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. 2023. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641* (2023).

[69] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. 2023. Dreamvideo: Composing your dream videos with customized subject and motion. *arXiv preprint arXiv:2312.04433* (2023).

[70] 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.

[71] 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).

[72] Yang Wu, Zhibin Liu, Hefeng Wu, and Liang Lin. 2023. Multi-object Video Generation from Single Frame Layouts. *arXiv preprint arXiv:2305.03983* (2023).

[73] Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. 2024. Parameter-Efficient Fine-Tuning for Pre-Trained Vision Models: A Survey. *arXiv preprint arXiv:2402.02242* (2024).

[74] 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).

[75] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16442–16453.

[76] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. 2024. Direct-a-Video: Customized Video Generation with User-Directed Camera Movement and Object Motion. *arXiv preprint arXiv:2402.03162* (2024).

[77] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. *arXiv preprint arXiv:2306.07954* (2023).

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

[79] Jaehong Yoon, Shoubin Yu, and Mohit Bansal. 2024. RACCooN: Remove, Add, and Change Video Content with Auto-Generated Narratives. *arXiv preprint arXiv:2405.18406* (2024).

[80] Jiwen Yu, Xiaodong Cun, Chenyang Qi, Yong Zhang, Xintao Wang, Ying Shan, and Jian Zhang. 2023. AnimateZero: Video Diffusion Models are Zero-Shot Image Animators. *arXiv preprint arXiv:2312.03793* (2023).

[81] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 69–85.

[82] Vladyslav Yushchenko, Nikita Araslanov, and Stefan Roth. 2019. Markov decision process for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[83] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[84] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. 2023. I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models. *arXiv preprint arXiv:2311.04145* (2023).

[85] Zhenfei Zhang and Ming-Ching Chang. 2023. Two-stage dual augmentation with clip for improved text-to-sketch synthesis. In *2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 1–6.

[86] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10668–10677.

[87] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. 2019. Image generation from layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8584–8593.

[88] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. 2018. Learning to forecast and refine residual motion for image-to-video generation. In *Proceedings of the European conference on computer vision (ECCV)*. 387–403.

[89] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. 2023. MotionDirector: Motion Customization of Text-to-Video Diffusion Models. *arXiv preprint arXiv:2310.08465* (2023).

[90] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. 2023. LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22490–22499.

[91] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018* (2022).

[92] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. 2024. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8806–8817.

[93] Junbao Zhuo, Xingyu Zhao, Shuhui Wang, Huimin Ma, and Qingming Huang. 2023. Synthesizing Videos from Images for Image-to-Video Adaptation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8294–8303.

Zero-Shot Controllable Image-to-Video Animation via Motion Decomposition

Supplementary Material

In this Appendix, we present the following:

- Additional details on grounding model implementation and human evaluation (Sec. 6).
- Additional details on the proposed benchmark for controllable image-to-video generation evaluation (Sec. 7).
- Attempt and observation on one-shot fine-tuning for better object consistency (Sec. 8).
- Additional qualitative results generated by the proposed IVA^0 and baseline models (Sec. 9).

6 MORE IMPLEMENTATION DETAILS

Grounding Model Implementation: We use GroundingDINO-B [41] with Swin-B backbone and pre-trained on COCO [40], O365 [54], GoldG [27], Cap4M [33], OpenImage [30], ODinW-35 [32], RefCOCO [81] for grounding detection in the evaluation pipeline. The box threshold was set to 0.35, and the text threshold was 0.25.
Human Evaluation: As discussed in Sec. 4.1, we designed a pipeline for human evaluation. Specifically, we first generate 100 example sets, where each example set contains 1 condition image and 3 videos generated by different methods (our IVA^0 and 2 baseline methods, in a shuffled order). Then, we ask 5 raters to rank among these 3 generated videos from the following aspects: 1) Controllability: Which method follows the object layout the best (We only consider whether the object is present in the given layout)? 2) Background Consistency: which method best maintains the background shown in the condition image? 3) Motion faithfulness: which method demonstrates the most plausible object out-of-place motion & object consistency? For each question, the raters are asked to select the best video. Our goal is to calculate the win rates of models along these 3 questions.

7 CONTROLLABLE I2V BENCHMARK BUILDING

Data Collection & Annotation: We gather 20 real-world images from public tracking/segmentation datasets [47] alongside 80 synthetic images created using the StableDiffusion-2.0 model [51]. 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 [8] that do not support layout input directly. We display the data—consisting of images, key boxes, and captions—in the format illustrated in Fig. 8.

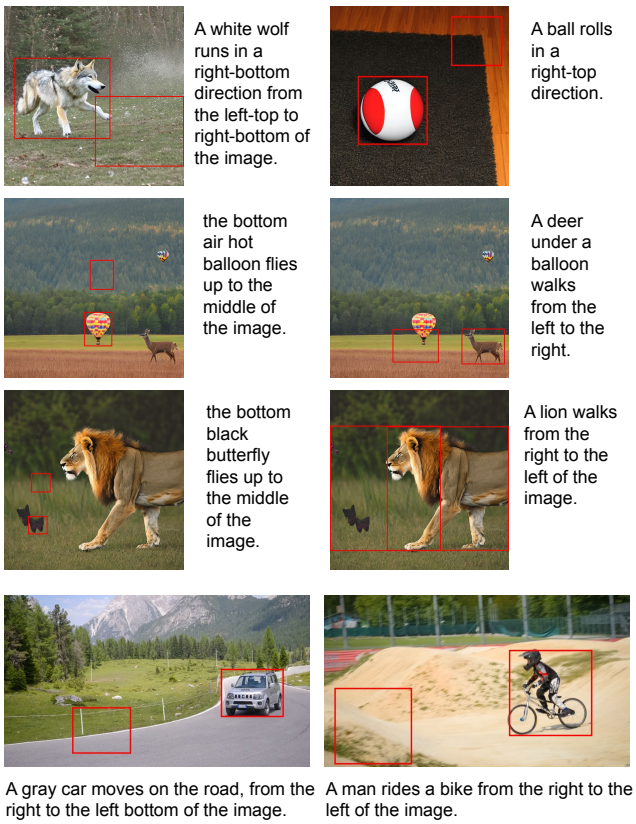

Figure 8: Data examples. We collect both real and generated images for our controllable Image-to-Video benchmark. The red boxes are manually annotated key boxes. We write motion-related captions for each image-layout pair for our baseline method. Best viewed in color.

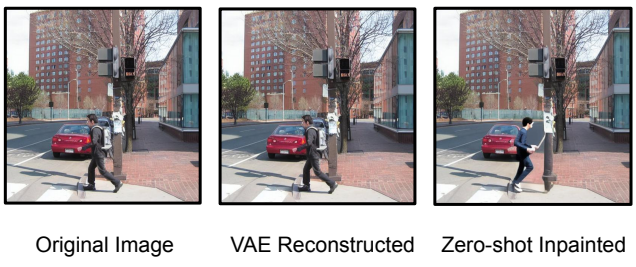

Figure 9: Comparison among the original image, VAE reconstructed image, and zero-shot inpainted image. We observe the inconsistency issue in the inpainted image due to the lack of low-level features.

Table 3: Comparison with recent video generation works.

Method	0-shot	I2V	Explicit Control	Inpainting-based
VideoComposer [66]	✗	✓	✓	✓
VideoCrafter [8]	✗	✓	✗	✗
AnimateZero [80]	✓	✓	✗	✗
MotionZero [59]	✓	✗	✗	✗
Motion-Zero [7]	✓	✗	✓	✗
Peekaboo [6]	✓	✗	✓	✗
IVA^0 (ours)	✓	✓	✓	✓

8 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. 9, 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 [45]. 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 [70, 71] 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 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 [57] 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.

More Detailed Comparison with Prior Works. We emphasize that our work provides new insight from image inpainting and motion decomposition views for a resource-friendly, user-controllable Image-to-Video (I2V) animation system. In Tab. 3, we compare our IVA^0 and highlight the difference with related recent controllable I2V/T2V works.

9 MORE QUALITATIVE RESULTS

In this section, we provide more qualitative results, Fig. 10 to Fig. 15 showcase additional generated video examples from our methods.

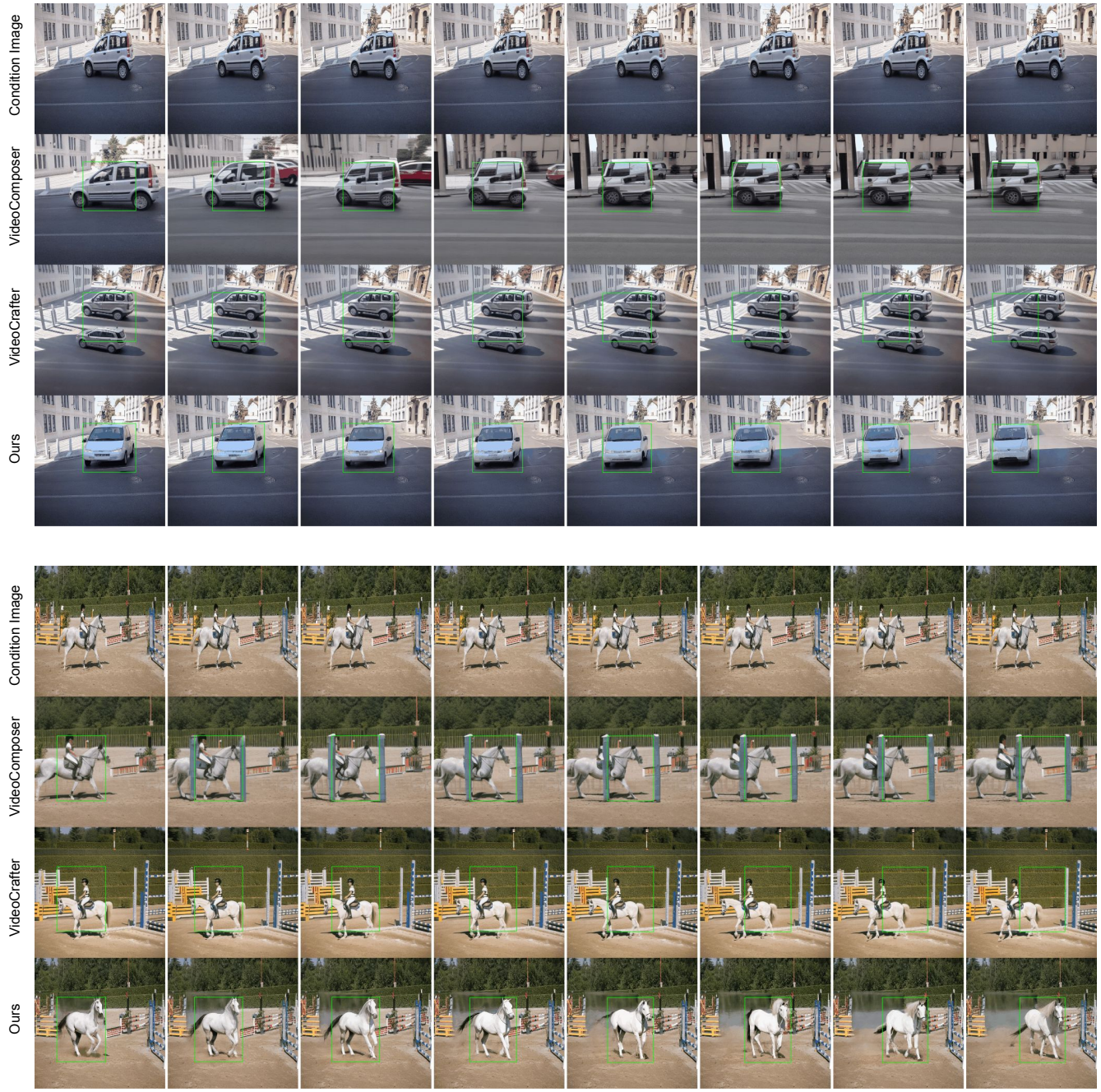

Figure 10: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.

Figure 11: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.

Figure 12: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.

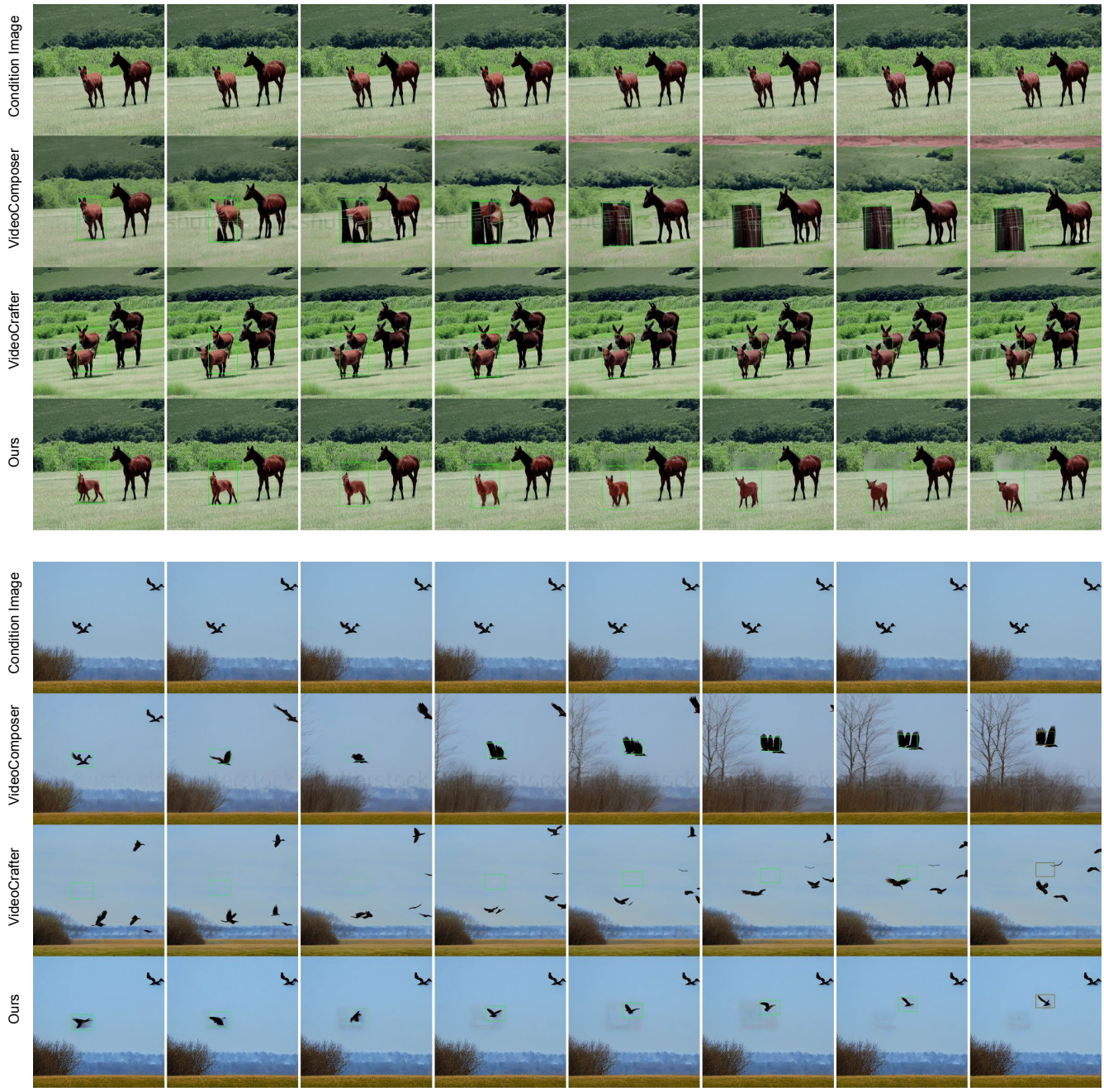

Figure 13: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.

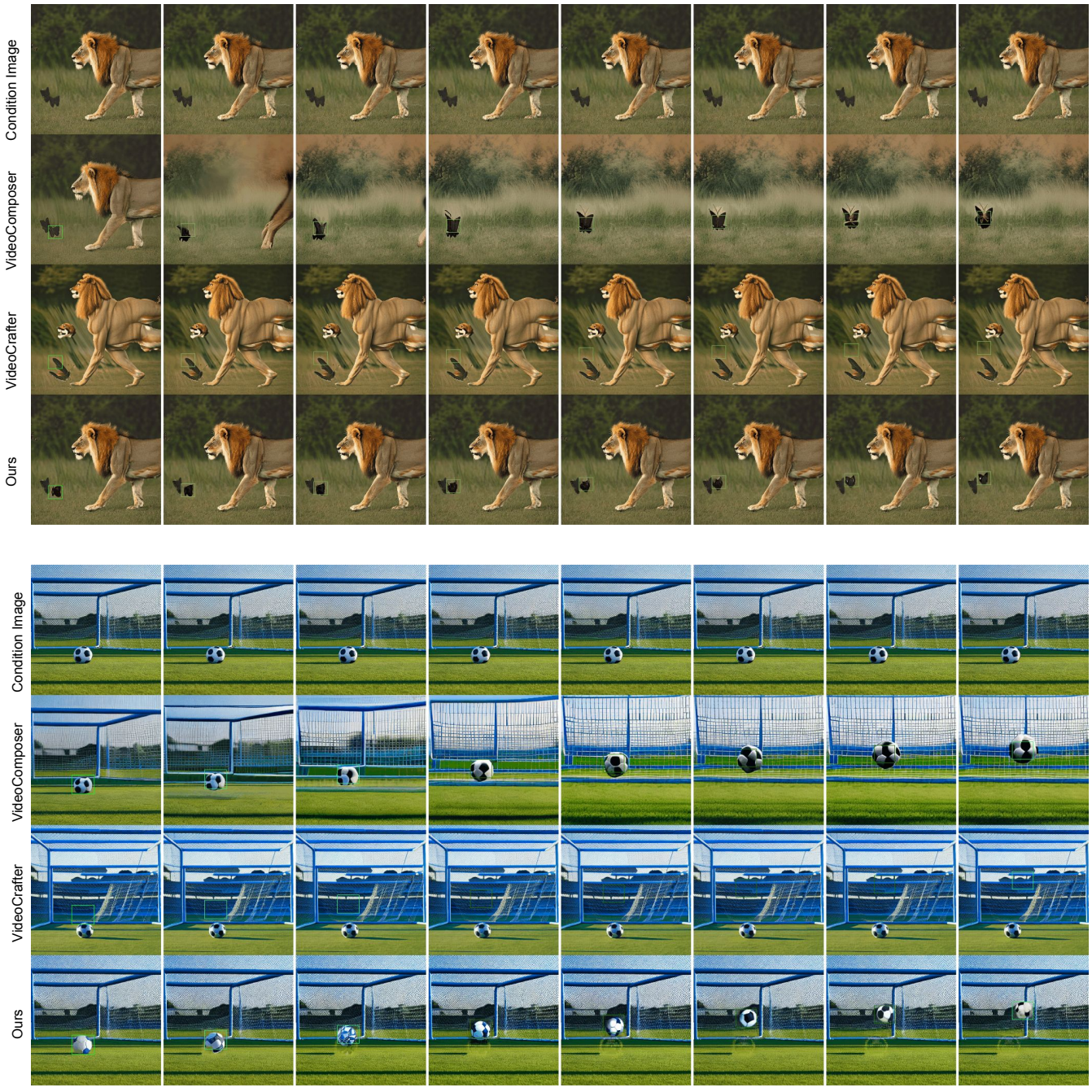

Figure 14: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.

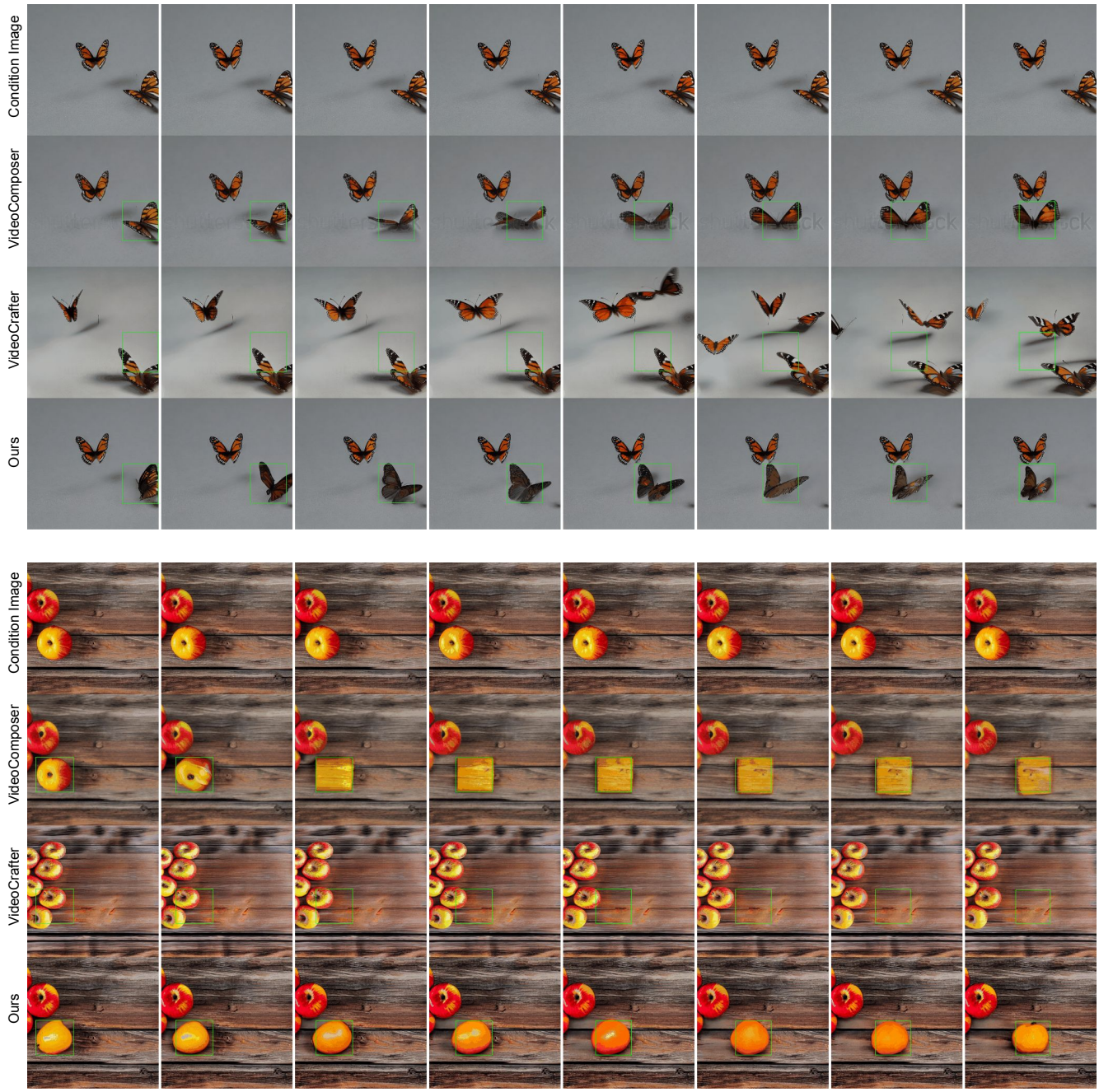

Figure 15: More generated examples comparison among proposed IVA^0 and baseline methods. The green boxes represent condition layout sequences. Best viewed in color.