IM-Zero: Instance-level Motion Controllable Video Generation in a Zero-shot Manner

Yuyang Huang[1]*, Yabo Chen[1]*, Li Ding[1], Xiaopeng Zhang[2]†, Wenrui Dai[1]†, Junni Zou[1]†, Hongkai Xiong[1], Qi Tian[2]

[1]Shanghai Jiao Tong University, Shanghai, China [2]Huawei Inc., Shenzhen, China

{huangyuyang, chenyabo, wangzhongren, daiwenrui, zoujunni, xionghongkai}@sjtu.edu.cn

zxphistory@gmail.com, tian.qi1@huawei.com

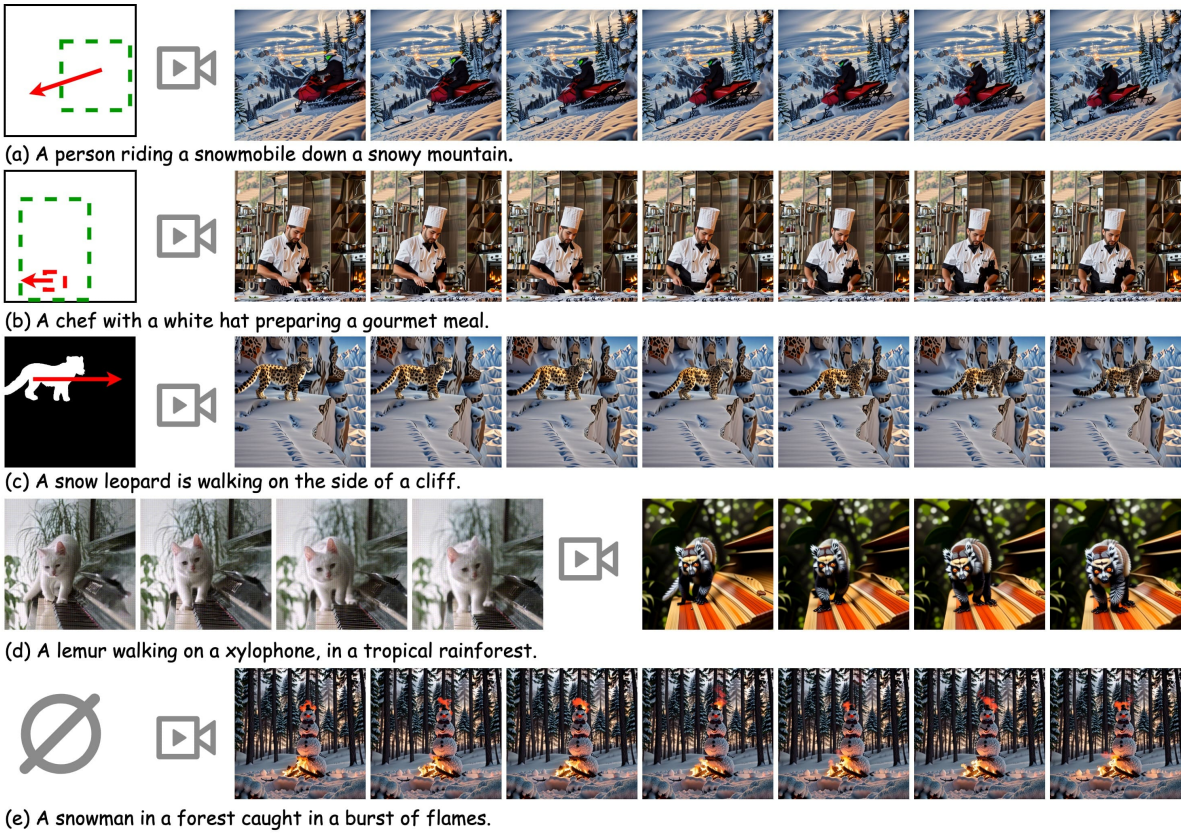

Figure 1. IM-Zero introduces a zero-shot framework for instance-level motion controllable video generation. IM-Zero allows users to: (a) Control overall instance motion. (b) Control instance subpart motion. (c) Control instance motion and specify instance shape via mask. (d) Perform motion transfer by reference video. (e) Perform high-quality text-to-video generation.

Abstract

Controllability of video generation has been recently concerned in addition to the quality of generated videos. The main challenge to controllable video generation is to synthesize videos based on user-specified instance spatial locations and movement trajectories. However, existing methods suffer from a dilemma between the resource consumption, generation quality, and user controllability. As an efficient alternative to prohibitive training-based video generation, existing zero-shot video generation methods cannot generate high-quality and motion-consistent videos under the control of layouts and movement trajectories. In this paper, we propose a novel zero-shot method named IM-Zero that ameliorates instance-level motion controllable video generation with enhanced control accuracy, motion consistency, and richness of details to address this prob-

*These authors contributed equally to this work.

†Corresponding authors: Xiaopeng Zhang; Wenrui Dai; Junni Zou.

lem. Specifically, we first present a motion generation *stage that extracts motion and textural guidance from keyframe candidates from pre-trained grounded text-to-image model to generate the desired coarse motion video. Subsequently, we develop a* video refinement *stage that injects the motion priors of pre-trained text-to-video models and detail priors of pre-trained text-to-image models into the latents of coarse motion videos to further enhance video motion consistency and richness of details. To our best knowledge, IM-Zero is the first to simultaneously achieve high-quality video generation and allow to control both layouts and movement trajectories in a zero-shot manner. Extensive experiments demonstrate that IM-Zero outperforms existing methods in terms of video quality, inter-frame consistency, and the alignment of location and trajectory. Furthermore, compared with existing methods, IM-Zero enjoys extra advantages of versatility in video generation, including motion control of subparts within instances, finer control of specifying instance shapes via masks, and more difficult tasks of motion transfer for customizing fine-grained motion patterns through reference videos and high-quality text-to-video generation.*

1. Introduction

Demands for enhanced controllability of text-to-video (T2V) generation are booming with the increasing quality of generated videos [3, 6, 7, 13–15, 22, 37, 38, 40, 62]. One of the critical problems lie in the precise control over the spatial layout and trajectories of instances within the generated videos. Existing methods fail to achieve fine-grained control by solely relying on text prompts to steer pre-trained T2V models, even employing fundamental T2V models with large-scale pre-training like Sora [4], KLING, and Doubao-PixelDance [53].

Existing methods usually extend ControlNet [55] for controllable video generation with frame-by-frame fine-grained control signals, including Canny, depth, and HED signals [9, 16, 36, 56]. To obtain dense control signals, these methods impose high demands on the input end and are limited in practical applications. One approach is to extract control signals from reference videos but it is hard to obtain reference videos perfectly matching the desired motion patterns. The other alternative requires manually crafting the dense control signals. However, it is also non-trivial to craft frame-by-frame control signals without expertise.

Recent attempts [20, 29, 32] use spatial locations (*i.e.*, bounding boxes) and motion trajectories of instances, along with corresponding prompts, to create desired videos. The locations and motion trajectories of instances are extracted from real videos and are encoded as additional control conditions for joint training with the video generation models. However, existing training-based methods can only control the movement trajectory [11, 41, 44, 52, 58, 65], since substantial data and training resources are required to simultaneously learn controlling layouts and movement trajectories for video generation.

In this paper, we develop a novel training-free approach to address this problem. Existing training-free methods are restricted in obtaining accurate location and movement effects by manipulating spatial attention maps and temporal attention maps to generate instances at specified locations [20, 29, 32]. These methods convert location signals in the pixel space provided by users to a latent space with reduced dimension for manipulating attention maps, and inevitably compromise the precision of location control and obtain inaccurate instance placement. Excessive constraints on the attention maps also disrupt the motion prior of video diffusion. Moreover, frame interaction is neglected during the manipulation, which also causes unreasonable motion effects and unnatural generated videos. Additionally, these methods also suffer from generating low-quality videos with choppy motion and insufficient details.

We find that both the inaccurate instance positioning and the unnatural motion stem from the limitations of attention map manipulation, which is a compromise solution to mitigate the lack of a clear understanding of instance concepts in vanilla latent diffusion models. It introduces constraints to enforce the instances to appear at designated positions. Contrary to attention map manipulation for video generation, instance-level knowledge is usually considered for image generation. Existing grounded T2I models like GLIGEN [25] and InstanceDiffusion [39] encapsulate extensive instance-level knowledge, but they are confined to generating images rather than consistent videos. Thus, we propose to generate coarse motion videos aligning with the user input and exhibit natural motion using instance-level priors in grounded T2I models, and further refine coarse motion videos by harnessing the motion and detail priors of T2V models for motion and detail injection in a zero-shot manner to boost the motion consistency and richness of details.

To this end, we propose a zero-shot and plug-and-play method named IM-Zero that allows to simultaneously control instance layouts and movement trajectories for video generation. IM-Zero consists of two stages, *i.e.*, a *motion generation* stage that acquires a coarse motion video, and a *video refinement* stage to enhance the quality of the coarse motion video. In the *motion generation* stage, spatial location signals (*e.g.*, bounding boxes and masks) are first augmented according to the input trajectory to obtain a sequence of layout control signals. A batch of keyframe candidates with texture consistency is then obtained by inflating and transforming the pre-trained grounded T2I model, and is used to extract motion guidance and textural guidance for generating the desired coarse motion video. In the *video refinement* stage, we encode the coarse motion video

into video latents, and inject the motion and detail priors extracted from the pre-trained T2V models into the video latents to improve motion consistency and enhance the richness of details, respectively.

Our contributions are summarized as below.

- We propose a novel framework named IM-Zero for zero-shot motion controllable video generation. IM-Zero supports various types of layout inputs and enables fine-grained motion control of the instances and their subparts.
- We present motion generation that generates keyframe candidates with the layout control capability of grounded T2I models to extract motion and textural guidance for coarse motion video generation.
- We develop zero-shot video refinement that injects the motion and detail priors of the pre-trained text-to-video model and text-to-image model into the latents of the coarse motion video to further enhance video motion consistency and richness of details.
- We further achieve motion transfer to customize complex motion patterns using reference videos and support high-quality text-to-video generation with only text input beyond instance layout and movement trajectory controls.

Extensive experiments demonstrate that IM-Zero outperforms existing methods in terms of video quality, inter-frame consistency, and the alignment of location and trajectory. Compared with existing methods, IM-Zero enjoys extra advantages of versatility in video generation, including motion control of subparts within instances and finer control with instance shapes specified by masked inputs. Furthermore, it can be adapted to motion transfer tasks to support the customized fine-grained motion patterns through reference videos and can also be adapted to high-quality text-to-video generation.

2. Related Work

Grounded Text-to-image Generation. Grounded text-to-image (T2I) models aim to generate images that align with input layout information. Some works specialize in modeling layout information of the bounding box type. Notable examples include GLIGEN [25], InstanceDiffusion [39], BoxDiff [46], MultiDiffusion [1], ReCo [49], Layout-Guidance [8], MIGC [63, 64], and TraDiffusion [43]. We believe that some training-based methods such as InstanceDiffusion [39] have enabled models to grasp the concept of instances. Their capability aligns to some extent with our objective of achieving location control and trajectory control in video generation. However, these methods are limited to generating single images and cannot produce consistent videos. Therefore, we aim to leverage Grounded T2I models to benefit controllable video generation tasks.

Location and Trajectory Control in Video Generation. Diffusion models have demonstrated exceptional capabilities in image generation [45, 61], video generation [56, 57], 3D generation [10, 27], and other visual tasks [18, 54, 60]. With the improvement in video generation quality [3, 6, 7, 13–15, 22, 37, 38, 40, 47, 62], there is a growing interest in achieving generation controllability, with a particular focus on controlling the spatial location and movement trajectories of instances. Currently, training-based methods can only achieve limited control over trajectory guidance, at the expense of requiring substantial amounts of data and computational resources, with typical examples being DragNUWA [52], DragAnything [44], DragVideo [11], Tora [58], TrackGo [65] and MotionCtrl [41]. In contrast, some training-free methods can concurrently control both spatial location and movement trajectories. For instance, Peekaboo [20] and TrailBlazer [29] enforce constraints on attention maps to position instances at specified locations. However, this approach can lead to issues such as inaccuracies in location and poor motion effects. FreeTraj [32] follows a similar attention map manipulation approach but introduces noise flow control. In addition to the aforementioned challenges, it also introduces difficulties in handling bounding boxes that undergo size changes. Therefore, we propose a zero-shot method that surpasses attention map manipulation to address the aforementioned challenges and enhance video quality and motion consistency.

3. Method

3.1. Preliminaries

Grounded Text-to-image Diffusion Models. Grounded text-to-image (T2I) diffusion models [25, 39, 63, 64] are capable of generating instances at user-specified locations by incorporating the layout input as conditions for joint training. InstanceDiffusion [39] supports four types of layout input: bounding box, mask, point, and scribble. It inserts learnable UniFusion blocks between the self-attention and cross-attention layers of the T2I model. UniFusion sample the instance layout as a 2D points $\mathbf{p}$ sequence and convert it to layout token by $\mathbf{g} = \mathrm{MLP}\left(\left[\tau_\theta\left(\mathbf{c}\right), \gamma\left(\mathbf{p}\right)\right]\right)$, where $\gamma(\cdot)$ is a Fourier mapping and $\mathbf{c}$ is the text prompt. Then the layout token is fused with the backbone visual tokens via a learnable masked self-attention layer.

Text-to-video Diffusion Models. Significant progress has also been made in Text-to-Video (T2V) diffusion models. While some T2V models are trained from scratch, the majority of them are trained based on existing T2I models. A typical example is AnimateDiff [13], which utilizes pre-trained T2I models as the spatial layer and trains additional motion modeling modules to reconstruct N frames clean video latents $z^{1:N}$ from i.i.d Gaussian noise conditioned on the text prompt c by $\mathcal{L}(\theta) = \mathbb{E}_{z_0^{1:N},\epsilon,c_t,t}\left[\|\epsilon - \hat{\epsilon}_\theta\left(z_t^{1:N}, c_t, t\right)\right]$.

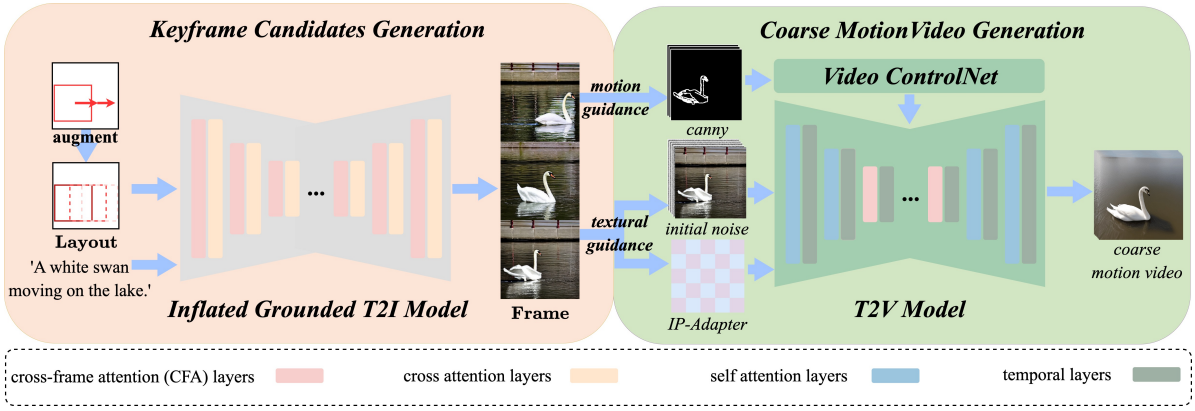

Figure 2. The Motion Generation Stage consists of two parts: Keyframe Candidates Generation and Coarse Motion Video Generation. In the first part, we augment the user input to obtain **Layout**, then generate keyframe candidates **Frame** with an inflated grounded T2I model. In the second part, we extract motion guidance and textural guidance from **Frame** to generate a coarse motion video.

3.2. Overall Framework

IM-Zero framework consists of two critical stages: Motion Generation Stage and Video Refinement Stage. In Section 3.1, we first briefly introduce the fundamentals of grounded T2I and T2V diffusion models. Then, we introduce the Motion Generation Stage in Section 3.3 as shown in Figure 2. This stage generates keyframe candidates and extracts guidance from them to generate a coarse motion video. Section 3.4 introduces the Video Refinement Stage as shown in Figure 3. This stage utilizes motion prior of T2V models to perform Motion Injection and utilizes the detail-generation capacity of T2I models to perform Detail Injection, further enhancing the visual quality of coarse motion video. Lastly, we also offer an option for motion transfer via reference video input, and we will explain how to easily modify IM-Zero in this case. As illustrated in Figure 1, IM-Zero achieves motion control for overall instances or subparts and allows users to further specify instance shapes as the grounded T2I model can generate keyframe candidates via masks. IM-Zero can perform motion transfer. Besides, IM-Zero allows high-quality T2V generation by refining any video via the Video Refinement Stage.

3.3. Motion Generation Stage

The core task involves presenting suitable motion patterns in the video, including two key aspects: the user-desired instance spatial location and the user-desired instance movement. Previous training-free methods achieved this task by manipulating attention maps but encountered issues such as inaccurate location, poor motion effects, and low video quality, as discussed in Section 1. IM-Zero aims to achieve the decoupling of instance spatial location and instance movement to enhance the performance of both aspects. For the former, we utilize a grounded T2I model [39] to generate instances at user-specified locations. For the latter, we depart from the previous approach of imposing excessive constraints on the attention map and instead efficiently leverage the motion priors of the T2V model.

Keyframe Candidates Generation. Considering the user input S layout signals and L trajectory segments, We denote the layout signals as $\mathbf{Layout} = \{s_i | i \sim [1, 2, ...S]\}$ (i.e., bounding boxes) and the movement trajectory as $\mathbf{Traj} = \{l_i | i \sim [1, 2, ..., L]\}$. We ensure that each l_i segment has a layout signal at both endpoints, for the endpoint without a layout signal, we select the closest s_i from **Layout** and duplicate it here, obtaining the augmented layout signals $\mathbf{Layout} = \{s_i | i \sim [1, 2, ..., S']\}$. These signals are then input into the grounded T2I model to generate keyframe candidates $\mathbf{Frame} = \{f_i | i \sim [1, 2, ..., S']\}$. Since generating each image separately can lead to significant inconsistencies, following [22], we inflate the grounded T2I model and replace each self-attention in UNet with cross-frame attention (CFA). The CFA for the i-th keyframe candidate f_i is defined as $\text{CFA}\,(Q^i, K^{1:N_{S'}}, V^{1:N_{S'}}) = \text{Softmax}(Q^i (K^1)^T / \sqrt{c}) V^1$ where Q,K and V denote the queries, keys and values. The first candidate provides key and value for all following ones, which can be interpreted from an image editing perspective as querying relevant content and textures from the first candidate to ensure consistency in textures [5, 22].

Coarse Motion Video Generation. After obtaining **Frame**, we utilize these candidates to generate a coarse motion video reflecting the user's input. To ensure smooth and natural instance motion, we have avoided the previous approach of directly enforcing excessive constraints on the attention map, which could disrupt the T2V motion priors. Instead, we extract two types of guidance from **Frame** to guide the T2V model. **(1) Motion Guidance.** To ensure the fidelity of the video to the user's input, encompassing spatial location and motion patterns, we extract foreground instances from **Frame** and convert them into Canny edge images. Subsequently, we employ ControlNet [12, 16, 55] to guide the generation of the coarse motion video. The rationale behind our approach is elucidated as follows: firstly,

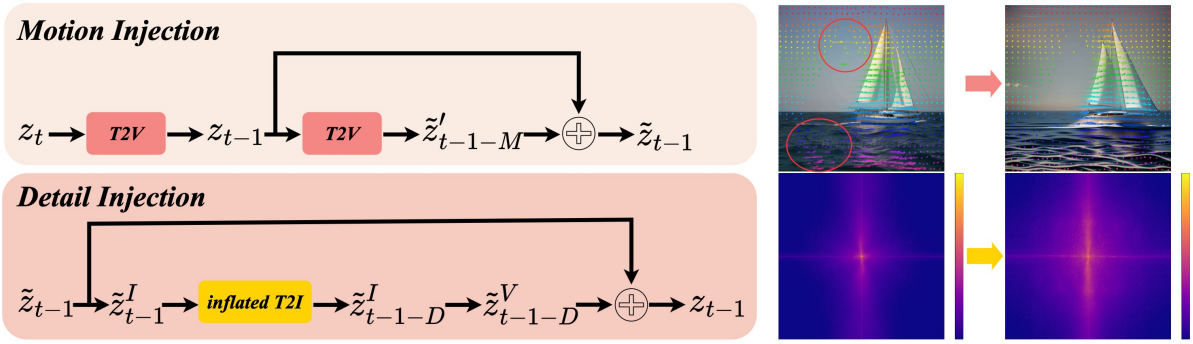

Figure 3. In Video Refinement Stage, we diffuse coarser motion video latents and perform motion injection and detail injection in the reverse process. In motion injection, we first obtain z_{t-1} by z_t, then denoise z_{t-1} for further M steps using a T2V model to $\tilde{z}'_{t-1-M}$ and combine them. We visualize the motion injection effect using CoTracker [21]. In detail injection, we first transform $\tilde{z}_{t-1}$ from T2V noise distribution to T2I noise distribution, then denoise it for further N steps with an inflated T2I model. The result is reversed to T2V noise distribution and combined with $\tilde{z}_{t-1}$. We visualize the detail injection effect using FFT analysis.

the separation of foreground from the background is essential because instances described in the user's input manifest as foreground instances in **Frame** [25, 39]. Although we apply operations to enhance consistency by CFA, background distortions and abrupt changes can still occur [22]. Secondly, transforming images into Canny edge representations not only aligns with the requirements of ControlNet but also further mitigates potential inconsistencies in instance textures within **Frame**. **(2) Textural Guidance.** Use Motion Guidance only may cause textural inconsistency, hence we further perform Textural Guidance. We diffuse the first keyframe candidate f_1 to form the initial noise of the T2V model. To help the guidance transmit from f_1 to subsequent frames, we devise the following strategies: firstly, we employ an IP-Adapter [51] to utilize f_1 as an image prompt. Moreover, we also apply CFA to the spatial layers of the T2V model.

3.4. Video Refinement Stage

We have observed that while coarse motion videos may align with user-desired inputs, unsmooth motion issues and suboptimal visual quality issues such as watermark and flickering exist. To attack these issues, we introduce a zero-shot method by more efficiently leveraging the motion priors inherent in existing T2V models and the detail-generation capacity of T2I models. As shown in Figure 3, we integrate a T2V model into the reverse process of video generation. We leverage T2V motion priors to inject motion into video latents, ensuring smoother and more natural motion. Subsequently, we also insert a T2I model, leveraging its detail generation capacity to enrich video detail richness.

Given a low-quality coarse motion video as described in Section 3.3, we start by encoding it using the encoder $\mathcal{E}$ of a vanilla T2V model (*i.e.*, AnimateDiff [13]) and adding Gaussian noise by timestep T to obtain noisy video latents z_T. For each denoising time step t, we first denoise z_t using the denoiser of the T2V model to obtain z_{t-1} by:

$$p_\theta\left(z_{t-1} \mid z_t\right)=\mathcal{N}\left(z_{t-1} ; z_t-\hat{\boldsymbol{\epsilon}}_\theta\left(z_t ; c, t\right),\left(1-\alpha_t^2\right) \mathbf{I}\right) \tag{1}$$

where c is the text prompt embedding. Then we sequentially perform motion injection and detail injection on z_{t-1}.

Motion Injection. We employ a vanilla T2V model (i.e., AnimateDiff [13]) for motion injection as shown in Figure 3. Specifically, we denoise z_{t-1} using its denoiser for further M steps to acquire $\tilde{z}'_{t-1-M}$, where each denoising step follows Equation (1). We observed a significant improvement in the temporal consistency of $\tilde{z}'_{t-1-M}$, resulting in very smooth motion. However, this enhancement comes at the cost of losing many detailed features. Therefore, we combine $\tilde{z}'_{t-1-M}$ with z_{t-1} to enhance motion smoothness while preserving details by:

$$\tilde{z}_{t-1}=\lambda_1 * z_{t-1}+\left(1-\lambda_1\right) * \tilde{z}'_{t-1-M} \tag{2}$$

where λ_1 is a coefficient controlling the fusion scale.

Detail Injection. We then employ an inflated T2I model for detail injection as shown in Figure 3. Due to the different noise distributions in the T2I and T2V models, a noise transformation is required. To convert video latents into image latents, we directly estimate clean latents [48] $\hat{z}_0$ by:

$$\hat{z}_0=\frac{\tilde{z}_t-\sqrt{1-\alpha_t} \epsilon_\theta\left(\tilde{z}_t, c, t\right)}{\sqrt{\alpha_t}} \tag{3}$$

and we then conduct DDIM inversion [30] to derive the corresponding noisy image latents $\tilde{z}^I_{t-1}$, the DDIM inversion process is defined by:

$$\tilde{z}^I_t=\sqrt{\alpha_t} \hat{z}_0+\sqrt{1-\alpha_t} \cdot \epsilon_\theta\left(\tilde{z}^I_{t-1}, c, t\right) \tag{4}$$

Subsequently, we denoise the noisy image latents with the inflated T2I model for further D steps to obtain $\tilde{z}^I_{t-1-D}$. The denoising process resembles Equation (1). As the T2I model excels in detail generation compared to the T2V

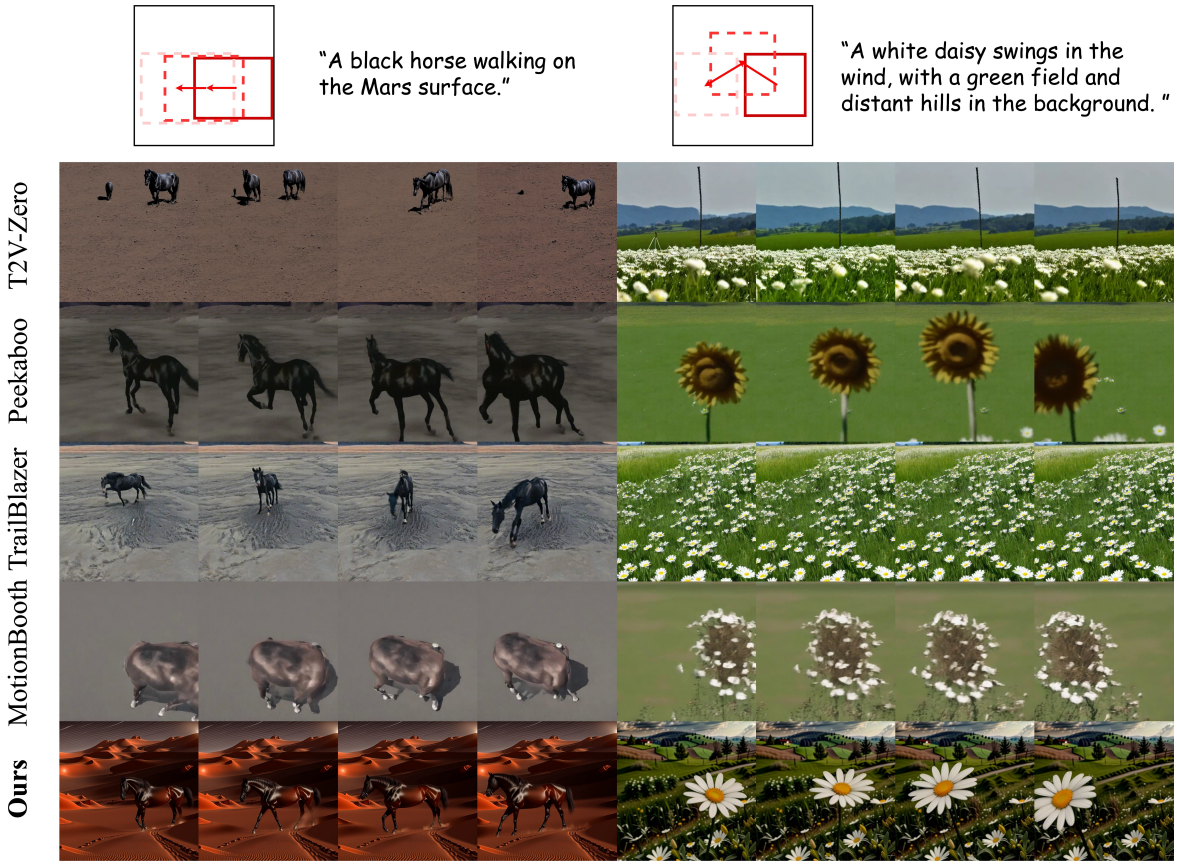

Figure 4. Qualitative Results of IM-Zero compared with other methods. Videos generated by IM-Zero outperform others in control accuracy, motion effects, and video quality.

model, we revert $\tilde{z}^I_{t-1-D}$ back to the noise distribution of the T2V model, combining the transformed latents with $\tilde{z}_{t-1}$ obtained from Equation (2) to integrate details. Specifically, in the reversed noise transformation process, we predict clean noise from $\tilde{z}_{t-1}$ using Equation (3) and utilize the DDPM forward process to obtain the corresponding noisy video latents $\tilde{z}^V_{t-1-D}$. Then we fuse them by:

$$z_{t-1} = \lambda_2 * \tilde{z}_{t-1} + (1 - \lambda_2) * \tilde{z}^V_{t-1-D} \tag{5}$$

where λ_2 controls the fusion scale and z_{t-1} is used for the next denoising timestep.

3.5. Adapt IM-Zero for Motion Transfer

For a more customized motion pattern provided by reference videos, IM-Zero can also be easily adapted for motion transfer tasks. Given a reference video and the target text prompt, we extract control signals (e.g., depth maps) from each frame of the reference video. Subsequently, we use these dense control signals and the target text prompt to guide the T2V model in generating a coarse motion video via ControlNet. Then, the coarse motion video undergoes video refinement stage as detailed in Section 3.4 to enhance motion consistency and video frame quality.

4. Experiment

4.1. Qualitative Results

Regarding the comparative methods, we primarily selected Peekaboo [20], TrailBlazer [29], and MotionBooth [42], which generate videos by specifying boxes and trajectories. Furthermore, to further illustrate the effectiveness of our approach, we also included T2V-Zero [22], which offers rough guidance by specifying translation vectors. To ensure a fair comparison, following the experimental settings of [29], all methods utilize the same layout input, trajectory input, and text prompt input. Additional results of controllable video generation are available in the supplementary material. We also compared with FreeTraj [32], but as it additionally requires inputting boxes with fixed sizes across frames to make noise flow, we place the comparison in the supplementary material.

As shown in Figure 4, our method significantly outperforms other approaches in terms of alignment with the input, motion effects, and video quality. In terms of alignment with the input, T2V-Zero exhibits almost no resemblance to the input, while the two methods based on attention map manipulation, Peekaboo, and TrailBlazer, sometimes roughly match the input location and trajectory but

Table 1. Quantitative results of IM-Zero compared with other methods on DAVIS-17 [31] dataset and GOT10k [17] dataset. IM-Zero achieves SOTA in all metrics, including metrics measuring video quality, frame consistency, and control accuracy.

Method	DAVIS-17					GOT10k				
	FVD (↓)	KID (↓)	CLIPSim (↑)	mIoU (↑)	CD (↓)	FVD (↓)	KID (↓)	CLIPSim (↑)	mIoU (↑)	CD (↓)
T2V-Zero [22]	5004.00	30.37	95.12	0.19	0.34	4335.69	24.72	95.41	0.12	0.32
Peekaboo [20]	4264.70	26.48	97.33	0.19	0.46	4960.49	19.05	98.68	0.11	0.34
TrailBlazer [29]	4593.37	26.04	98.10	0.21	0.29	4695.18	16.18	97.80	0.13	0.40
MotionBooth [42]	4192.19	26.85	96.64	0.22	0.36	4311.71	16.05	97.47	0.14	0.42
IM-Zero (Ours)	**4070.01**	**25.75**	**98.91**	**0.27**	**0.26**	**4187.27**	**15.92**	**99.04**	**0.21**	**0.27**

Table 2. Ablation on different components in Motion Generation Stage.

Method	FVD (↓)	KID (↓)	CLIPSim (↑)	mIoU (↑)	CD (↓)
w/o 1st CFA	4320.07	26.84	98.84	0.26	0.29
w/o 2nd CFA	4528.09	28.08	98.78	0.21	0.31
w/o init noise	4786.17	26.69	98.82	0.26	0.31
w/o IP-Adapter	4549.94	27.75	98.87	0.26	0.28
IM-Zero (Ours)	**4070.01**	**25.75**	**98.91**	**0.27**	**0.26**

still lack precision. Regarding motion effects, Peekaboo and TrailBlazer generate videos with unnatural motion, resulting in strange discontinuities between frames. Concerning video quality, Peekaboo and TrailBlazer exhibit low image quality, lack detail, and introduce artifacts. We also provide a video version of Figure 4 in the supplementary material.

4.2. Quantitative Results

Experiment Settings. For datasets, we utilize the training set and validation set of the DAVIS-17 dataset [31] and the test set of the GOT10k dataset [17]. We use BLIP-2 [24] to generate text prompts and GroundingDINO [28] to extract bounding boxes from the first and last frame as inputs of all the methods. After eliminating videos where no bounding boxes are extracted, we randomly select 50 videos and 70 videos from the rest videos of DAVIS-17 [31] and GOT10k [17]. For our method, in the first stage we use InstanceDiffusion [39] as grounded T2I model and SparseCtrl [12] as ControlNet, in the second stage we use Stable Diffusion V1.5 [34] as T2I model, on both stage we use AnimateDiff [13] as T2V model. For the evaluation metrics, we employ mean Intersection over Union (mIoU) and Centroid Distance (CD) to assess control alignment, Fréchet Video Distance (FVD) [35] and Kernel Inception Distance (KID) [2] to evaluate video quality, and CLIP similarities (CLIPSim) [33] to assess frame-to-frame consistency. Detailed settings are available in the supplementary material.

Metrics Evaluation Results. We conducted evaluations on the aforementioned metrics on the DAVIS-17 dataset [31] and the GOT10k dataset [17]. As shown in Table 1, our method achieved state-of-the-art (SOTA) performance compared to several other methods. Firstly, in terms of the alignment between the generated videos and user inputs, our method demonstrates a significant lead. Compared to the previous best methods, our approach achieved an improvement of 28.6% and 61.5% in mIoU, and reductions of 10.4% and 15.6% in CD on the two datasets, indicating that our method offers more accurate location control and motion trajectory control compared to previous methods based on attention map manipulation. Secondly, regarding frame-to-frame consistency in the videos, thanks to our enhanced consistency techniques and motion injection methods, our approach also reaches the SOTA performance. Moreover, in terms of video quality, our method significantly outperforms others, with FVD reduced to 4070.01 and 4187.27, and KID reduced to 25.75 and 15.92, respectively.

Table 3. Motion transfer results compared with other methods.

Method	Motion Fidelity (↑)	Imaging Quality (↑)
MotionClone [26]	74.5	0.72
MOFT [23]	51.2	0.71
MotionDirector [59]	**75.5**	0.68
IM-Zero (Ours)	63.7	**0.73**

Table 4. T2V results compared with UNet baseline.

Method	Imaging Quality (↑)	CLIPSim (↑)
AnimateDiff [13]	0.67	99.05
IM-Zero (Ours)	**0.74**	**99.73**

4.3. Versatile Capacity Analysis

Our method possesses versatile capacity as shown in Figure 5. First, our method can control the subparts of instances, achieving effects such as head turning and hand

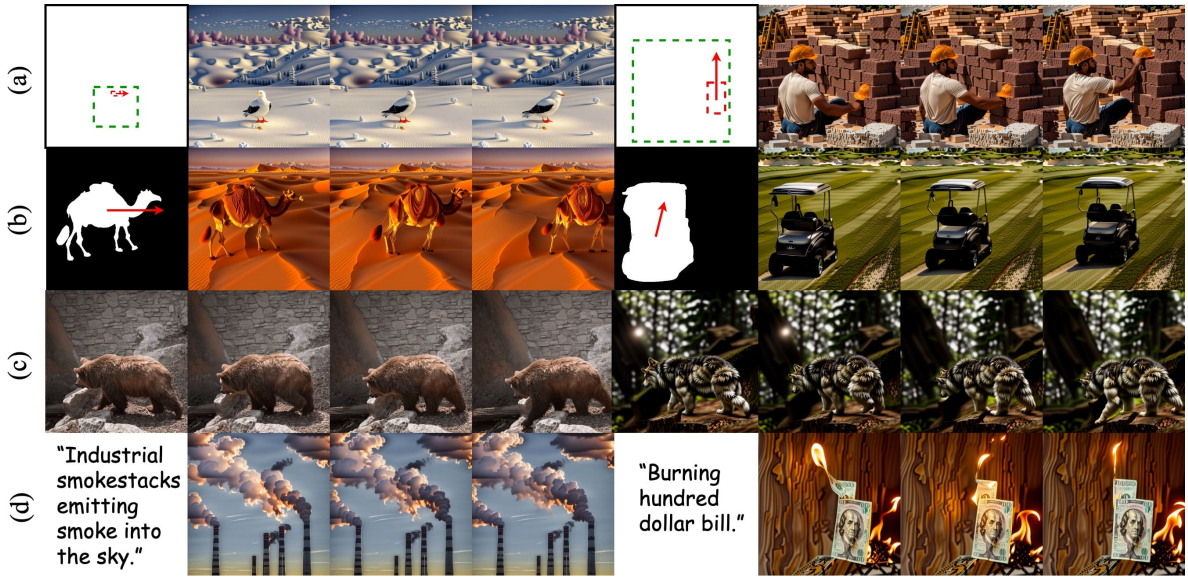

Figure 5. Besides box-level overall instance motion control, IM-Zero can: (a) Perform instance subpart control. (b) Specify instance shape by mask. (c) Perform motion transfer from the reference video(left) to the target video(right). (d) Perform high-quality T2V generation.

movement. Second, besides bounding boxes, we also support specifying instance shapes more precisely using masks. Third, for cases where users provide a reference video, our method can perform motion transfer. Finally, users can directly provide text inputs for high-quality T2V generation. For motion transfer and T2V generation, we compare IM-Zero with other methods as follows.

Motion Transfer. We compare with MOFT [23], MotionClone [26], and MotionDirector [59] using the open-sourced data from MotionClone [26]. For quantitative results, we use Motion Fidelity [50] and Imaging Quality [19] as metrics following MOFT [23]. Table 3 shows that our method is competitive in Motion Fidelity [50] and is the best in Imaging Quality [19]. Qualitative results and other details are provided in the supplementary material.

T2V Generation. As we use UNet-based models, we compare T2V with UNet baseline AnimateDiff [13]. For quantitative results, we use Imaging Quality [19] and CLIP-Sim [33] as metrics. As there are no ground-truth videos, we generate videos using 50 prompts randomly generated by GPT. Table 4 shows that our method is better in Imaging Quality [19] and CLIPSim [33]. Refer to the supplementary material for qualitative results and other details.

Other Capacities. The proposed method also supports multi-instance motion control, and the decomposition of our pipeline allows autoregressive usage of the second stage to enhance temporal consistency, as elaborated in the supplementary material.

4.4. Ablation Studies

Ablation on different components in Motion Generation Stage. We ablate on different components in Section 3.3. As depicted in Table 2, these methods improve the quality of the final videos and also enhance the alignment with user inputs. Refer to the supplementary material for details.

Ablation on different components in Video Refinement Stage. We also conduct ablation studies on Video Refinement Stage. Refer to the supplementary material for details.

5. Conclusion and Limitation

We introduce IM-Zero, a novel training-free framework for instance-level motion controllable video generation using diffusion models. We leverage grounded text-to-image models to generate keyframe candidates and extract motion guidance and textural guidance from them to generate videos that align with user inputs. To further enhance video motion and quality, we also introduce a zero-shot method for motion injection and detail injection. Additionally, our method can control the subpart motion of instances, use masks to specify instance shapes, perform motion transfer from reference videos, and directly support high-quality text-to-video generation.

Limitation. The zero-shot features impose certain limitations on our method. Firstly, as zero-shot methods depend on pre-trained model weights, the videos generated by our method are inherently constrained by the pre-trained models employed. For example, the alignment with layout input is influenced by the grounded T2I model and video ControlNet, the video quality is affected by the T2V model, which can be improved by using more advanced pre-trained models. Secondly, the temporal consistency of videos generated by zero-shot methods is inherently weaker compared to training-based methods, which can be improved by incorporating some tuning components to our method.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62125109, Grant 62431017, Grant U24A20251, Grant 62320106003, Grant 62371288, Grant 62401357, Grant 62401366, Grant 62301299, Grant 62120106007, and in part by the Program of Shanghai Science and Technology Innovation Project under Grant 24BC3200800.

References

[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1737–1752, 2023. 3

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *The Sixth International Conference on Learning Representations*, 2018. 7

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 3

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Blog, 2024. 2

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 4

[6] 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. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2, 3

[7] 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. 2, 3

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

[9] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-A-Video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 2

[10] Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascade-Zero123: One image to highly consistent 3D with self-prompted nearby views. In *Proceedings of the 18th European Conference on Computer Vision*, pages 311–330, 2024. 3

[11] Yufan Deng, Ruida Wang, Yuhao Zhang, Yu-Wing Tai, and Chi-Keung Tang. DragVideo: Interactive drag-style video editing. In *Proceedings of the 18th European Conference on Computer Vision*, pages 183–199, 2024. 2, 3

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. SparseCtrl: Adding sparse controls to text-to-video diffusion models. In *Proceedings of the 18th European Conference on Computer Vision*, pages 330–348, 2024. 4, 7

[13] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 5, 7, 8

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

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems 35*, pages 8633–8646, 2022. 2, 3

[16] Zhihao Hu and Dong Xu. VideoControlNet: A motion-guided video-to-video translation framework by using diffusion model with ControlNet. *arXiv preprint arXiv:2307.14073*, 2023. 2, 4

[17] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2019. 7

[18] Yuyang Huang, Yabo Chen, Yuchen Liu, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. DomainFusion: Generalizing to unseen domains with latent diffusion models. In *Proceedings of the 18th European Conference on Computer Vision*, pages 480–498, 2024. 3

[19] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 8

[20] 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*, pages 8079–8088, 2024. 2, 3, 6, 7

[21] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *Proceedings of the 18th European Conference on Computer Vision*, pages 18–35, 2024. 5

[22] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 2, 3, 4, 5, 6, 7

[23] Karlo Koledić, Igor Cvišić, Ivan Marković, and Ivan Petrović. MOFT: Monocular odometry based on deep depth and careful feature selection and tracking. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6175–6181. IEEE, 2023. 7, 8

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742, 2023. 7

[25] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2, 3, 5

[26] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. MotionClone: Training-free motion cloning for controllable video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 7, 8

[27] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 3

[28] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the 18th European Conference on Computer Vision*, pages 38–55, 2024. 7

[29] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. TrailBlazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024*, number 97, 2024. 2, 3, 6, 7

[30] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 5

[31] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 7

[32] 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. 2, 3, 6

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 7, 8

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 7

[35] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[36] Cong Wang, Jiaxi Gu, Panwen Hu, Haoyu Zhao, Yuanfan Guo, Jianhua Han, Hang Xu, and Xiaodan Liang. EasyControl: Transfer ControlNet to video diffusion for controllable generation and interpolation. *arXiv preprint arXiv:2408.13005*, 2024. 2

[37] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 3

[38] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. MagicVideo-V2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024. 2, 3

[39] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. InstanceDiffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6232–6242, 2024. 2, 3, 4, 5, 7

[40] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. LaVie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 2024. 2, 3

[41] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3

[42] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. MotionBooth: Motion-aware customized text-to-video generation. In *Advances in Neural Information Processing Systems 37*, pages 34322–34348, 2024. 6, 7

[43] Mingrui Wu, Oucheng Huang, Jiayi Ji, Jiale Li, Xinyue Cai, Huafeng Kuang, Jianzhuang Liu, Xiaoshuai Sun, and Rongrong Ji. TraDiffusion: Trajectory-based training-free image generation. *arXiv preprint arXiv:2408.09739*, 2024. 3

[44] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. DragAnything: Motion control for anything using entity representation. In *Proceedings of the 18th*

European Conference on Computer Vision, pages 331–348, 2024. 2, 3

[45] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[46] 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*, pages 7452–7461, 2023. 3

[47] Yichao Yan, Zanwei Zhou, Zi Wang, Jingnan Gao, and Xiaokang Yang. DialogueNeRF: Towards realistic avatar face-to-face conversation video generation. *Visual Intelligence*, 2 (1):24, 2024. 3

[48] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22873–22882, 2023. 5

[49] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. ReCo: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023. 3

[50] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8476, 2024. 8

[51] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 5

[52] 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, 3

[53] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8850–8860, 2024. 2

[54] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements DINO for zero-shot semantic correspondence. In *Advances in Neural Information Processing Systems 36*, pages 45533–45547, 2023. 3

[55] 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, 4

[56] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. ControlVideo: Training-free controllable text-to-video generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3

[57] Yabo Zhang, Yuxiang Wei, Xianhui Lin, Zheng Hui, Peiran Ren, Xuansong Xie, Xiangyang Ji, and Wangmeng Zuo. VideoElevator: elevating video generation quality with versatile text-to-image diffusion models. *arXiv preprint arXiv:2403.05438*, 2024. 3

[58] Zhenghao Zhang, Junchao Liao, Menghao Li, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. *arXiv preprint arXiv:2407.21705*, 2024. 2, 3

[59] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. MotionDirector: Motion customization of text-to-video diffusion models. In *Proceedings of the 18th European Conference on Computer Vision*, pages 273–290, 2024. 7, 8

[60] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5706–5716, 2023. 3

[61] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. CogView3: Finer and faster text-to-image generation via relay diffusion. In *Proceedings of the 18th European Conference on Computer Vision*, pages 1–22, 2024. 3

[62] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. MagicVideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2, 3

[63] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. MIGC: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828, 2024. 3

[64] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. MIGC++: Advanced multi-instance generation controller for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1714–1728, 2025. 3

[65] Haitao Zhou, Chuang Wang, Rui Nie, Jinxiao Lin, Dongdong Yu, Qian Yu, and Changhu Wang. TrackGo: A flexible and efficient method for controllable video generation. *arXiv preprint arXiv:2408.11475*, 2024. 2, 3