TrackDiffusion: Tracklet-Conditioned Video Generation via Diffusion Models

Pengxiang Li[1*] Kai Chen[2*] Zhili Liu[2,3*] Ruiyuan Gao[4] Lanqing Hong[3]
Dit-Yan Yeung[2] Huchuan Lu[1] Xu Jia[1†]

[1]Dalian University of Technology [2]Hong Kong University of Science and Technology
[3]Huawei Noah's Ark Lab [4]The Chinese University of Hong Kong

Abstract

Despite remarkable achievements in video synthesis, achieving granular control over complex dynamics, such as nuanced movement among multiple interacting objects, still presents a significant hurdle for dynamic world modeling, compounded by the necessity to manage appearance and disappearance, drastic scale changes, and ensure consistency for instances across frames. These challenges hinder the development of video generation that can faithfully mimic real-world complexity, limiting utility for applications requiring high-level realism and controllability, including advanced scene simulation and training of perception systems. To address that, we propose ***TrackDiffusion****, a novel video generation framework affording fine-grained trajectory-conditioned motion control via diffusion models, which facilitates the precise manipulation of the object trajectories and interactions, overcoming the prevalent limitation of scale and continuity disruptions. A pivotal component of TrackDiffusion is the instance enhancer, which explicitly ensures inter-frame consistency of multiple objects, a critical factor overlooked in the current literature. Moreover, we demonstrate that generated video sequences by our TrackDiffusion can be used as training data for visual perception models. To the best of our knowledge, this is the first work to apply video diffusion models with tracklet conditions and demonstrate that generated frames can be beneficial for improving the performance of object trackers.* [1]

1. Introduction

Benefiting from the development of the diffusion models, video generation has achieved breakthroughs, particularly in text-to-video (T2V) generation models [16, 41]. The utilization of diffusion models and large-scale text-video pairs markedly expanded the ability to generate diverse video content [2, 5, 11, 17, 26, 46] enabling a more nuanced interpretation of textual prompts and translating them into dynamic, visually compelling narratives. Although textual descriptions provide a friendly interactive manner for image generation, it is not easy for them to impose fine control over generated content. Several control signals have been employed to generate images with more flexibility and higher quality, such as control signals from segmentation, content edges [47], and object boxes [6, 10, 20] to specify object or image layout. Considering video's nature of continuity and temporal dynamics, textual descriptions also can not provide sufficient information to guarantee highly realistic details, even Sora [30] may fail in case of spontaneous appearances of objects [3]. Approaches like MOFA-Video [29] and MotionClone [23] have been proposed to address these issues by introducing additional motion control signals, allowing for enhanced control over the generated content. However, fine-grained motion control could contribute much to high-quality video generation. Such fine-grained control not only enhances visual quality but also has the potential to enable applications like perception model training, animated storytelling, and advanced user interfaces.

While fine-grained motion control is a natural interaction for video generation, it is still under-explored, especially for diffusion-based video generation models. Despite the progress [40, 45] in the field, existing generative models often fail to maintain instance-level consistency across frames critical for reproducing the complex temporal dynamics found in natural settings. Consequently, they struggle to capture the dynamic interplay among multiple objects, especially in complex scenarios marked by occlusion, overlapping objects, and unpredictable rapid movements as depicted in Fig. 1.

In this work, we introduce *TrackDiffusion*, a novel framework specifically designed to fill this lacuna. Integrating with video diffusion models, *TrackDiffusion* enables fine-grained motion control of generated contents with object boxes. Specifically, we first introduce instance-aware location tokens for each object, which embed identity infor-

[1]* Equal contribution. †Corresponding author.

(a) Prompt: A black and tan dog and its three puppies play **in the center** of a sunny grass field.

(b) Prompt: A man in a red shirt **maintains a central position** in the frame while riding a mountain bike along the ground.

(c) Prompt: A train is entering the station, passing another stationary train slowly, **approaching the camera**.

Figure 1. **Qualitative comparison on the trajectory-conditioned motion control**. ModelScope [37] does not support controls other than text. In contrast, the generation results of *TrackDiffusion* are more consistent with the input prompts.

mation of boxes into boxes across frames, and are helpful in addressing the object occlusion and re-occurrence. Besides, one distinctive component of our framework is the *instance enhancer* module. This simple yet effective component provides inter-frame consistency of objects, ensuring remarkable instance-level consistency. Finally, gated cross-attention is employed to seamlessly integrate the box conditions into a pretrained video diffusion model such that the huge amount of computation for training from scratch could be avoided.

Our extensive experiments demonstrate that *TrackDiffusion* surpasses prior methods in the quality of the generated video data. Furthermore, ablation studies confirm the necessity of introducing instance-aware location tokens and instance enhancer for achieving these results. Our experiment also shows that generated videos by *TrackDiffusion*, as augmented data, could benefit tracking tasks and bring further improvement on the performance of tracking models.

The main contributions of this work contain three parts:

1. We present the **very first known** application of DMs to generate continuous video sequences directly from the tracklets, a methodological innovation that transcends the capabilities of existing video generative models.

2. A novel component of our framework, the *instance enhancer*, is proposed to provide consistent inter-frame object identity, even in challenging conditions such as occlusion and rapid movement.

3. Our experimental results demonstrate that by incorporating tracklet constraints, the quality of the videos improves substantially, and the track average precision (TrackAP) score of the object tracker, which assesses the alignment between the given boxes and the generated objects, experiences a significant boost, underscoring the efficacy of motion control.

2. Related Work

2.1. Layout-to-Image Generation

Layout-to-image (L2I) generation, focusing on converting high-level graphical layouts into photorealistic images, has witnessed considerable advancements. GLIGEN [20] enhances the pre-trained diffusion models with gated self-attention layers for improved layout control, while LayoutDiffuse [8] employs novel layout attention modules tailored

for bounding boxes. Instead, GeoDiffusion [6] enables various geometric controls directly via text prompts to support object detection [12, 19] data generation, which is further extended for concept removal by Gemo-Erasing [24], and 3D geometric control by MagicDrive [10].

2.2. Text-to-Video Generation

Text-to-video (T2V) generation, following the successful trajectory of the text-to-image (T2I) generation, has achieved significant advancements. Most of the T2V methodologies [13, 14, 49] tend to focus on depicting the continuous or repetitive actions from textual prompts, rather than capturing the dynamics of multiple, changing actions or events. However, these methods generally lack the ability to generate complex transitions and diverse event sequences. On the other hand, recent works such as LVD [21] employ large language models to create dynamic scene layouts for video diffusion, concentrating on text-driven layout generation. VideoComposer [38] enables conditions such as sketches, depth maps, and motion vectors. VideoDirectorGPT [22] and DriveDreamer [39] have furthered multi-scene video generation, showcasing advancements in the field.

2.3. Point-based editing

To enable fine-grained editing, several works have been proposed to perform point-based editing. DragNUWA [45] suggests video generation conditioned on an initial image, provided trajectories, and text prompts. DragDiffusion [33] studies drag-based editing with diffusion models. MotionCtrl [40] propose a unified motion controller that can use either the camera poses and object trajectories to control the motion of generated videos. However, these methods did not provide a precise way to define objects, which makes it difficult to select and manipulate larger or composite objects within an image. Additionally, trajectories fail to account for an object's shape and size, both of which are essential for representing changes in movements.

3. Method

In this section, we first introduce the latent diffusion model (LDM [32]), on which our method is based, in Section 3.1. Then, we introduce the *TrackDiffusion* pipeline, the *instance-aware location tokens* and *temporal instance enhancer*, in Section 3.2. We also present our methods for enhancing instance consistency across frames, particularly in the video clips with noticeable spatial changes. An overview of *TrackDiffusion* is shown in Fig. 2.

3.1. Preliminary: Latent Diffusion Models (LDM)

Recent advancements in image synthesis have been significantly driven by LDM. These models excel by focusing on the distribution within the latent space of images, marking a notable leap in performance in this domain. The LDM comprises two main components: an autoencoder and a diffusion model.

The autoencoder is responsible for compressing and reconstructing images, utilizing an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. Specifically, the encoder projects an image x into a lower-dimensional latent space z, followed by the decoder reconstructing the original image from this latent representation. The reconstruction process yields an image $\tilde{x} = \mathcal{D}(z)$, approximating the original image x. Given that the data distribution $z_0 \sim q(z_0)$ is progressively corrupted by Gaussian noise over T steps, this process follows a variance schedule denoted by $\beta_1, \ldots, \beta_T$:

$$q(z_t \mid z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbb{I}\right), \quad t = 1, \ldots, T \tag{1}$$

with a U-Net, $\epsilon_\theta(z_t; t)$, trained to predict this added noise using a loss function:

$$L(\theta) = \mathbb{E}_{t \sim \mathcal{U}(1,T), \epsilon_t \sim \mathcal{N}(0,1)} \left[\left\| \epsilon_t - \epsilon_\theta(z_t; t, \mathbf{y}) \right\|^2 \right], \tag{2}$$

where x_t is the noisy sample of x_0 at timestep t. The condition $\mathbf{y}$ can be $\varnothing$ (unconditional generation), text [32] or images [15], etc.

3.2. Tracklet-Conditioned Video Generation

3.2.1 Overview.

Our method, *TrackDiffusion*, introduces an innovative approach to video generation from tracklets, addressing the challenges of instance consistency and spatial-temporal coherence in complex video sequences. The methodological backbone of *TrackDiffusion* consists of four pivotal components: *Instance-Aware Location Tokens*, *Temporal Instance Enhancer*, *Motion Extractor*, and *Gated Cross-Attention*. Together, these components form a synergistic framework that not only captures the intricacies of individual frames but also preserves the natural flow and continuity of multi-object interactions across a video sequence.

3.2.2 Task Definition.

The primary objective of *TrackDiffusion* is to generate high-fidelity video sequences from tracklets, where a tracklet refers to a sequence of object bounding boxes across frames, coupled with their respective category information. Formally, given a set of tracklets $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$, where each tracklet τ_i corresponds to an object instance across T frames, our task is to generate a video sequence V that accurately represents the motion and appearance of these instances. Each tracklet τ_i is defined as $\tau_i = \{(b_{i,1}, c_{i,1}), (b_{i,2}, c_{i,2}), \ldots, (b_{i,T}, c_{i,T})\}$, where $b_{i,t}$ denotes the bounding box coordinates of the i-th instance in frame t, and $c_{i,t}$ represents the category of the instance. The

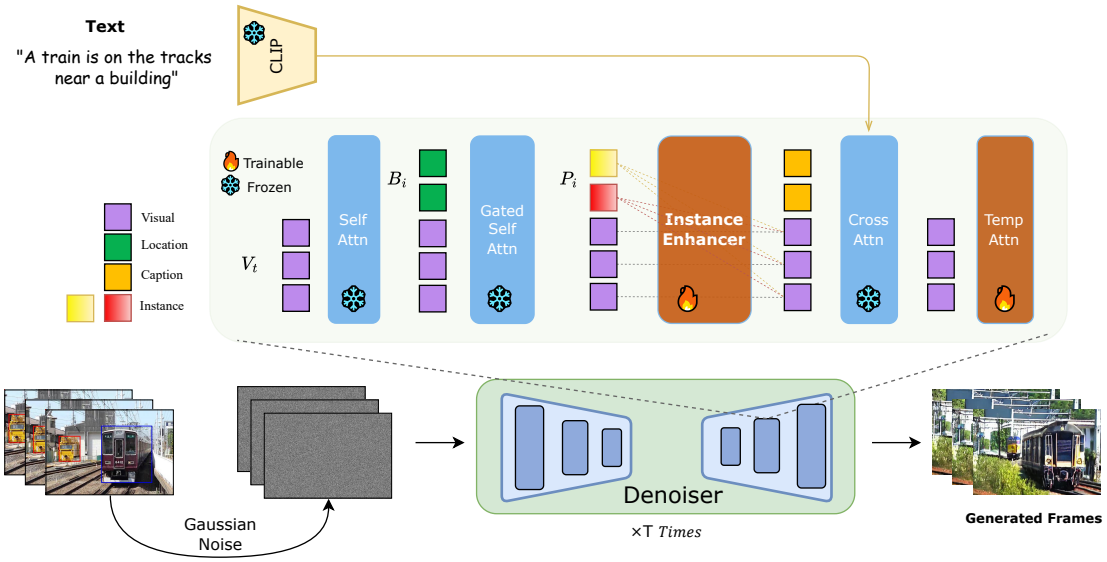

Figure 2. **Model architecture of *TrackDiffusion***. The framework generates video frames based on the provided tracklets and employs the *Instance Enhancer* to reinforce the temporal consistency of foreground instance. A new gated cross-attention layer is inserted to take in the new instance information.

generated video V is a sequence of frames $\{v_1, v_2, \ldots, v_T\}$, where each frame v_t is a synthesis of the instances as per their tracklet descriptions at time t.

3.2.3 Instance-Aware Location Tokens.

To fully leverage the condition of object layouts, the coordinates of a 2D object box $b_{i,t}$ in a frame are projected into the embedding space similarly to the positional encoding in GLIGEN [20]. This projection, $B_{i,t} = \text{Fourier}(b_{i,t})$, is then concatenated with the box's category embedding and transformed into the conditioning representation $H_{i,t}$, as:

$$H_{i,t} = \text{MLP}([c_{i,t}, B_{i,t}]), \tag{3}$$

where $c_{i,t}$ denotes the category embedding of the i-th bounding box computed with the CLIP model [31], and Fourier$(\cdot)$ refers to Fourier embedding [28].

To encourage consistency of instances across frames, we further complement the layout condition representation with instance identity information. In this way, the trajectories of various instances in the sequence could be distinguished, and the continuity of instances across frames would be reinforced. The enhanced instance-aware location token is represented as $H'_{i,t} = H_{i,t} + e_i$, with e_i representing a learnable token for the instance denoted by the i-th box in the frame.

Subsequently, a gated self-attention [20] is employed to impose conditional layout information into visual features as shown below:

$$V_t = V_t + \tanh(\beta) \cdot \text{TS}(\text{SelfAttn}([V_t, H'_{:,t}])), \tag{4}$$

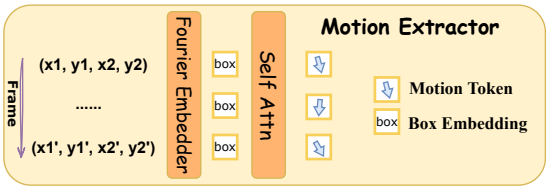

Figure 3. **Illustration of Motion Extractor.** We perform self-attention on the Fourier-encoded bounding box coordinates to integrate motion information.

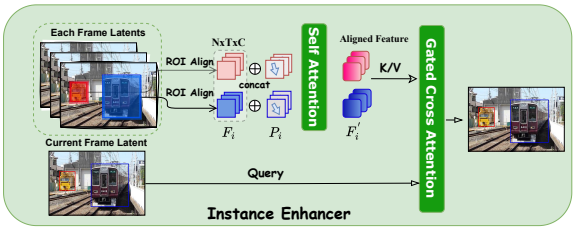

Figure 4. **Illustration of Temporal Instance Enhancer**. Each instance's latent features, after *ROI Align*, are concatenated with motion tokens and then processed through self-attention layers. We demonstrate the specific enhance operation using the yellow instance as an example.

where V_t denotes the visual feature tokens in frame t, $H'_{:,t}$ represents the enhanced location tokens for all boxes at time t, β is a trainable parameter, and TS$(\cdot)$ is the token selection operation focusing only on visual tokens at each time frame.

We refer readers to check [20] for more details.

3.2.4 Temporal Instance Enhancer.

In the context of tracklet conditioned video generation, a significant challenge is to maintain consistency of generated instances across frames, especially when instances have large spatial displacement in the sequence. Existing works employ temporal attention to encourage temporal consistency over frames where attention works at each position along time. However, when there is dramatic motion with the object or the camera, the attention computed on the same position would not work very well (see Fig. 5). To address this challenge, we propose a novel temporal attention called *Temporal Instance Enhancer* where attention is computed on the same instance instead of position, as explained in the following.

I. Instance-Level Feature Extraction. Let F_i represent the multi-dimensional feature tensor for the i-th instance, constructed by temporally concatenating features extracted from each frame t using the provided bounding box through ROIAlign. This process aligns the features of the same instance across all temporal frames to the same spatial size:

$$F_i = \bigoplus_{t=1}^{T} \text{ROIAlign}(V_t, B_{t,i}), \tag{5}$$

where $\bigoplus$ denotes the concatenation operation, T is the number of frames, and V_t represents the latent of the current U-Net block, as shown in Fig. 4. We simultaneously use a box consistent with the latent shape to perform ROIAlign, representing the background feature tensor.

II. Motion Extractor. Furthermore, we aim to integrate trajectory representation into instance-level features to enhance its capacity for capturing temporal dynamics. The trajectory representation is computed by applying self-attention to a sequence of box embeddings, each representing the same instance across successive frames, as shown in Eqn. 6 and Fig. 3. This method not only captures the spatial coordinates of the instance at each time step (*i.e.*, the location information) but also, by analyzing these sequences, discerns the movement of the instance over time (*i.e.*, the motion information). The inclusion of self-attention enables the model to effectively track objects smoothly through occlusions and interactions by inferring the continuity of object presence and movement. It is the model's ability to detect subtle shifts in position and temporal dependencies through self-attention that facilitates the accurate modeling of motion trajectories, as it comprehends the dynamic changes in an instance's location from frame to frame.

Similar to temporal attention, self-attention is then applied to the feature representation for each instance along time, as shown in Eqn. 7 and Fig. 4.

$$P_i = \text{SelfAttn}(B_{1,i}, B_{2,i}, \ldots, B_{T,i}) \tag{6}$$

$$F_i' = \text{SelfAttn}(F_i \oplus P_i), \tag{7}$$

where F_i' represents the enhanced features for i-th instance, which plays an important role in promoting accurate and consistent video generation.

III. Instance-Level Feature Integration. To integrate the enhanced instance features in video generation, we borrow the idea from GLIGEN and design a gated cross-attention layer which is inserted after the gated self-attention layer. This layer can make full use of temporally enhanced instance features for consistent video generation. The visual feature tokens from the gated self-attention layer are represented as $V = [v_1, \ldots, v_M]$, where M denotes the total number of tokens in the flattened latent. Then the gated cross-attention layer could be simply formulated as below:

$$V = V + \text{CrossAttn}([V, F_i']), \tag{8}$$

where F_i' denotes the enhanced instance features. This insertion ensures that the visual tokens in the LDM framework are now additionally informed by the aggregated instance features, thereby maintaining the consistency of instance appearance across different frames.

4. Experiments

Given that the traditional text-to-video (T2V) datasets (*e.g.*, MSR-VTT [42] and UCF101 [34]) typically do not contain bounding box annotations, we turn to adopt tracking datasets that offer precise tracklet annotations. This allows us to appraise our model's performance in generating text-to-video content within the multi-object tracking scenarios, where tracklet-conditioned video synthesis is essential. Our primary quantitative evaluations are conducted using a version of our model that extends the ModelScopeT2V [37] framework.

4.1. Implementation Details

4.1.1 Dataset.

Our experiments purposefully utilize both the YTVIS2021 [43] and the MOT-17 [27] datasets with a unified objective: to validate effectiveness of our approach in maintaining the consistency across multi-object generation. The YTVIS2021 dataset, serving as a cornerstone in VIS literature, includes 2,985 training videos with high-quality bounding box annotations from 40 semantical classes. On the other hand, the MOT-17 is a prominent

Method	Output Res.	Frames	Latency (s/frame)	YoutubeVIS		
				FVD↓	TrackAP↑	TrackAP$_{50}$↑
Oracle*	-	-	-	-	45.4	64.1
CogVideo(Eng.) [16]	160×160	16	-	1384	-	-
LVDM [13]	256×256	16	-	1011	-	-
ZeroScopeT2V [37]	576×320	16	-	750	-	-
Show-1 [46]	576×320	16	-	704	-	-
VideoCrafter [4]	256×256	16	-	690	-	-
ModelScopeT2V [37]	256×256	16	0.19	786	2.7	9.4
ControlVideo [48]	512×512	16	0.26	760	10.5	19.1
VideoComposer [38]	256×256	16	0.29	738	19.8	30.6
Vanilla	256×256	16	0.26	603	36.0	56.2
TrackDiffusion	256×256	16	0.28	605	39.4(+3.4)	62.0(+5.8)
TrackDiffusion	480×320	16	0.32	**548**	**44.7**(+8.7)	**68.0**(+11.8)
TrackDiffusion(SVD [1])**	512×320	25	0.77	**312**	**45.1**(+9.1)	**69.3**(+13.1)

Table 1. **Comparison of generation fidelity on YoutubeVIS datasets.** Vanilla is our customized baseline, similar to DriveDreamer [39], fine-tuned on the YTVIS2021 dataset. *: represents the real image *Oracle* baseline. **: We use Stable Video Diffusion as the base generation model instead of ModelScope.

Prompt: Two race cars speed along the track, a car passes by another car closely.

Prompt: Two white planes are flying, preparing to land at the airport.

Figure 5. **Qualitative comparison** among different methods. The input text prompt is shown on the bottom side of the figure.

dataset in MOT research, encompassing over 10,000 frames that focus on the pedestrian tracking. For the YTVIS2021 dataset, due to the absence of annotations for the validation set, we have randomly selected 160 videos from the training set to serve as our validation set. Regarding MOT-17, we divide the MOT-17 training dataset in half, using one half for training and the other half for validation, following common practice. To compensate the absence of captions

in the YTVIS2021 and MOT-17 datasets, we utilized the VideoBLIP [18] model to generate captions for each video clip.

4.1.2 Evaluation metrics.

In our evaluation, we utilize the captions and box annotations from the validation sets of YTVIS. We adopt *FVD* [35] to evaluate video quality. To evaluate the grounding accuracy, which is the correspondence between the input bounding box and the generated instance, we use the *Tracking Average Precision* (*TrackAP* [44]). This involves using pretrained QDTrack [9] model to track objects in the generated videos, which are then compared with the ground truth tracklets. It's important to note that since previous text-to-video methods do not support incorporating the box annotations as input, it is not equitable to compare them on this metric. Therefore, we limit our comparison to report FVD scores for reference. FVD scores on more datasets (*e.g.*, UCF101) are provided in Appendix A.

4.1.3 Baseline.

We introduce a baseline model, **Vanilla**, in Tab. 1, conceptually motivated by the DriveDreamer [39] framework. This model represents our specialized adaptation for layout-conditioned video generation, incorporating key principles from the DriveDreamer methodology, yet distinctly engineered to align with the realm of layout-conditioned video generation on YoutubeVIS.

4.1.4 Details.

We implement our approach based on the Diffusers [36] code base for ModelScopeT2V. Our training methodology comprises two stages. In **Stage 1**, we focus on *single-image layout controllability* by removing all temporal layers and employing gated self-attention. **Stage 2** *extends the approach to video data*, introducing *temporal attention*, *temporal convolution*, and an *instance enhancer* for video-level control. We trained the *Stage 1* on the corresponding training set for 60,000 steps, the *Stage 2* for 50,000 steps. The training process was carried out on 8 NVIDIA Tesla 80G-A800 GPUs. During generation, frames are sampled using the DPM-Solver [25] scheduler for 50 steps with the classifier-free guidance (CFG) set as 5.0.

4.2. Comparison with Existing T2V Methods

In this section, we conduct a comprehensive evaluation of the proposed *TrackDiffusion* with regard to video quality and trajectory controllability, which demands a realistic representation of objects while being consistent with geometric layouts.

4.2.1 Video Quality.

Tab. 1 compares our *TrackDiffusion* model with a range of recent video synthesis methods on the YTVIS dataset. Our approach demonstrates a significant advancement in the field, evidenced by competitive FVD scores which serve as a metric for video quality. Specifically, *TrackDiffusion* at a resolution of 256×256 achieves an FVD score of 605, showcasing its effectiveness in synthesizing high-fidelity videos. Notably, an enhanced version of TrackDiffusion, operating at a higher resolution of 480×320, further improves the FVD score to 548. This performance surpasses several contemporary models, including Vanilla, VideoCrafter, and others, underscoring the significance of the fine-grained control and consistency mechanisms implemented in *TrackDiffusion*. Notably, while the standard version aligns with the resolution of many counterparts, the high-resolution variant sets a new benchmark in the field. This is primarily because, at higher resolutions, the instance enhancer can extract cleaner instance features, laying a solid foundation for improved video quality. These results underscore the capability of *TrackDiffusion* to not only maintain but also enhance the quality of video synthesis, even when scaling to higher resolutions. This indicates that the model can effectively handle increased complexity and detail, a critical factor for realistic video generation.

Furthermore, as demonstrated in Fig. 5, we present the generation results of VideoComposer, Vanilla, and *TrackDiffusion*. For VideoComposer, we control video generation using depth maps. Vanilla struggles to maintain consistency in the appearance of instances. The color of the racing car changes in the first example, and in the second example, the object is even lost. Despite the control signals provided by the depth maps, VideoComposer's granularity remains coarse. Due to the absence of box-level control, the quality of the generated objects is somewhat inferior, and it fails to accurately produce high-quality videos that align with the user intended motion control.

4.2.2 Trajectory Controllability.

Tab. 1 also showcases the trajectory control precision of various video synthesis models, with a particular focus on the TrackAP metric from the YoutubeVIS dataset. *TrackDiffusion*, in both its standard and high-resolution variants, demonstrates a superior ability to precisely control trajectory. The standard 256×256 resolution version of *TrackDiffusion* achieves a TrackAP score of 39.4, which is an improvement of 3.4 points over the Vanilla model. When the resolution is increased to 480×320, *TrackDiffusion*'s performance further improves, reaching a TrackAP score of 44.7, which marks an 8.7 point increase. It's important to note that TrackAP scores in our experiments do not equate to the success rate in trajectory control. For comparison, we

Setting	Instance-Tokens	Instance-Enhancer	Motion-Extractor	FVD↓	TrackAP↑	TrackAP$_{50}$↑
(a)				741	34.2	60.4
(b)	✓			765	35.0(+0.8)	61.7(+1.3)
(c)	✓	✓		**729**	**38.7**(+4.5)	**64.3**(+3.9)
(d)	✓	✓	✓	**698**	**38.9**(+4.7)	**64.6**(+4.2)

Table 2. **Ablation of the instance-aware location tokens, the temporal instance enhancer and the motion extractor.** Defaults are marked in gray.

tested tracker performance using real data, as indicated by the Oracle results, which we consider an approximate upper bound for TrackAP scores. The closer our TrackAP scores are to the Oracle benchmark, the better the generated data's fidelity. Therefore, we should concentrate on the comparative difference in TrackAP scores between methods rather than their absolute values.

4.3. Ablation Study

To ascertain the effectiveness of our proposed design moudles, we conducted an ablation study focusing on crucial components of the model, such as the instance tokens and the instance enhancer. These evaluations were carried out using the YTVIS validation set, as discussed in Sec. 4.1.

4.3.1 Setup.

We conduct ablation studies primarily focusing on fidelity and report the results for FVD and TrackAP metrics. To balance training duration and the adverse effects of lower resolutions on the tracker, our experiments in this section generate videos at a resolution of 384×256. We employ Mask2Former [7] to evaluate TrackAP, mitigating the impact of data noise.

4.3.2 Effectiveness of Consistency Module.

The ablation study in Tab. 2 examines the impact of incorporating instance embeddings on instance consistency, using FVD and TrackAP as metrics. In Setting (a), without instance embeddings, the model shows baseline performance. However, introducing instance embeddings in Setting (b) leads to a slight increase in FVD (from 741 to 765), suggesting a minor trade-off in video quality. Despite this, TrackAP improves by 0.8, indicating enhanced instance consistency. This trade-off may be attributed to the additional parameters and not yet fully optimized training. Notably, when provided with sufficient training time, as shown in Tab. 1, instance embedding does not negatively impact FVD. The most significant improvements are seen in Setting (c), where both instance embeddings and temporal instance enhancer are employed, further confirming the effectiveness of these features in improving instance consistency.

4.4. Effectiveness of Motion Extractor.

We manually curate a subset of 23 videos from the validation set, ensuring that each video encompasses instances of target overlap or re-occurrence to varying degrees. We aim to demonstrate the effectiveness of the proposed motion information extraction through experiments on this subset. Results indicate a marked improvement when the motion extractor is employed. Specifically, the inclusion of the motion extractor yields a decrease in FVD to **774** from 793 and an increase in TrackAP by **2.0** points, achieving a score of 36.5. This enhancement is also reflected in the TrackAP$_{50}$ score, which sees an increase of **2.5** points, reaching 59.0. These results corroborate the efficacy of the motion extractor in our model, signifying its essential role in capturing and maintaining coherent motion trajectories in complex video scenes.

5. Conclusion

In conclusion, our work presents *TrackDiffusion*, a novel approach to generating continuous video sequences from tracklets, effectively utilizing diffusion models for video synthesis in the context of multi-object tracking. Our model introduces innovative mechanisms, including Instance-Aware Location Tokens and Temporal Instance Enhancer, which together facilitate the generation of high-quality and temporally consistent video sequences. The experimental results also show the potential of *TrackDiffusion* in enhancing the training of perception models, thereby marking a significant step forward in the realm of synthetic video data generation. Future work will focus on addressing the outlined limitations, further improving the model's generalization capabilities, and exploring its applicability in a broader range of real-world scenarios.

Acknowledgments. We gratefully acknowledge supports of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research. This research has been made possible by funding support from the Research Grants Council of Hong Kong through the Research Impact Fund project R6003-21. The research was partially supported by the National Natural Science Foundation of China, (grants No. 62472065, U23B2010,62106036).

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 6

[2] 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 *CVPR*, 2023. 1

[3] Bloomberg. Openai's sora video generator is impressive, but not ready for prime time. https://www.bloomberg.com/news/newsletters/2024-02-22/openai-s-sora-video-generator-is-impressive-but-not-ready-for-prime-time, 2024. Accessed: 2024-03-07. 1

[4] 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. 6

[5] 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. *arXiv preprint arXiv:2401.09047*, 2024. 1

[6] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *arXiv preprint arXiv:2306.04607*, 2023. 1, 3

[7] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 8

[8] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023. 2

[9] Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. In *PAMI*, 2023. 7

[10] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 1, 3

[11] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 1

[12] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. Soda10m: Towards large-scale object detection benchmark for autonomous driving. *arXiv preprint arXiv:2106.11118*, 2021. 3

[13] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2023. 3, 6

[14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[15] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. In *JMLR*, 2022. 3

[16] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 1, 6

[17] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 1

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 7

[19] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022. 3

[20] 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. *arXiv preprint arXiv:2301.07093*, 2023. 1, 2, 4, 5

[21] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 3

[22] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. 3

[23] 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. *arXiv preprint arXiv:2406.05338*, 2024. 1

[24] Zhili Liu, Kai Chen, Yifan Zhang, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, and James Kwok. Geom-erasing: Geometry-driven removal of implicit concept in diffusion models. *arXiv preprint arXiv:2310.05873*, 2023. 3

[25] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 7

[26] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023. 1

[27] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 5

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Communications of the ACM*, 2021. 4

[29] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222*, 2024. 1

[30] OpenAI. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators, 2024. Accessed: 2024-03-07. 1

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[33] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8849, 2024. 3

[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[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] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 7

[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, 5, 6

[38] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *NeurIPS*, 2023. 3, 6

[39] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023. 3, 6, 7

[40] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023. 1, 3

[41] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022. 1

[42] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 5

[43] Linjie Yang, Yuchen Fan, Yang Fu, and Ning Xu. The 3rd large-scale video object segmentation challenge - video instance segmentation track, 2021. 5

[44] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 7

[45] 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. 1, 3

[46] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 1, 6

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1

[48] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 6

[49] 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. 3