EMERGING TRACKING FROM VIDEO DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We find video diffusion models, renowned for their generative capabilities, surprisingly excel at pixel-level object tracking without any explicit training for this task. We introduce a simple and effective method to extract motion representations from video diffusion models, achieving state-of-the-art tracking results. Our approach enables the tracking of identical objects, overcoming limitations of previous methods reliant on intra-frame appearance correspondence. Visualizations and empirical results show that our approach outperforms recent self-supervised tracking methods, including the state-of-the-art, by up to 6 points. Our work demonstrates video generative models can learn intrinsic temporal dynamics of video, and excel in tracking tasks beyond original video synthesis.

1 INTRODUCTION

"What I cannot create, I do not understand."

– *Richard Feynman*

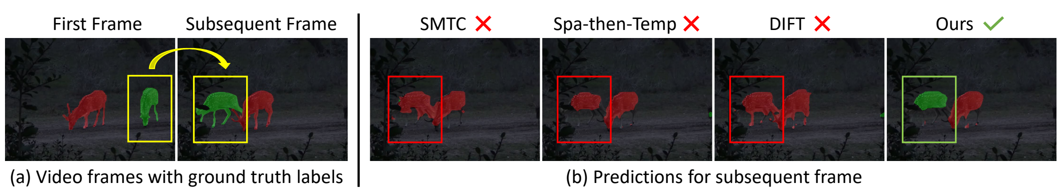
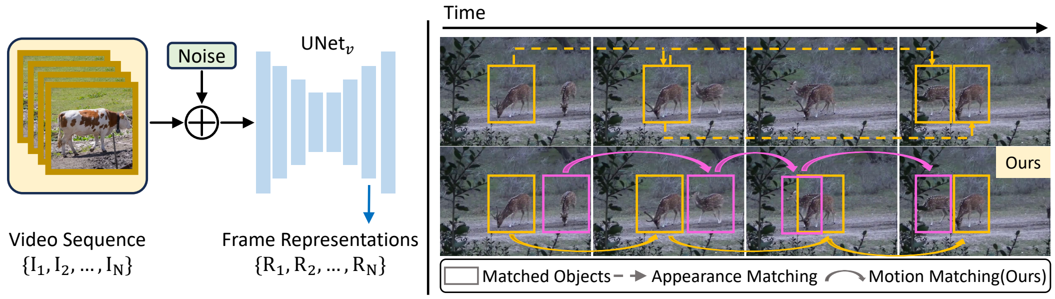


Figure 1: **Predictions from video label propagation task.** State-of-the-art models fail to find the correct temporal correspondence when multiple objects look similar in a video, such as SMTC (Qian et al., 2023), Spa-then-Temp (Li & Liu, 2023), and DIFT (Tang et al., 2023). For instance, the deer with green segmentation map labels in (a) are mislabeled as red by existing models, as highlighted by the red boxes in (b). By introducing latent representations from pretrained video diffusion models, our method captures temporal motions and correctly identifies the deer, highlighted by the green box in (b). Our work significantly improves tracking performance across various scenarios.

The ability of temporal relational reasoning over time (Yi et al., 2019) is crucial for visual intelligence. Rather than performing simple appearance correspondence, people often rely on temporal relational reasoning to track moving objects in complex situations (Yi et al., 2019; Gerstenberg et al., 2015; Ullman, 2015). For example, given the two moving deer in Figure 1(a), we can easily reason and track different deer even after they change their relative positions.

Learning video representations for temporal correspondence is essential for tasks like video object segmentation (Caron et al., 2021). Appearance-based correspondence methods have been used for tracking (Wang et al., 2021; Hu et al., 2022), including the recent state-of-the-art DIFT (Tang et al., 2023) that uses latent representations from image diffusion models (Rombach et al., 2021; Dhariwal & Nichol, 2021). Some research also integrates temporal information in model training (Wang et al., 2019; Jabri et al., 2020). However, existing methods often have low accuracy because they fail to capture temporal context in complex scenarios, see Figure 1(b), where state-of-the-art models (Qian et al., 2023; Li & Liu, 2023; Tang et al., 2023) fail to differentiate between two deer.

In this paper, we demonstrate that representations from video diffusion models can improve tracking across various scenarios, including those with multiple objects of similar appearance. Video diffusion



(a) Representations from video diffusion model (b) Our work improves tracking by introducing temporal motions

Figure 2: **Framework.** Our work focuses on the video label propagation task, which uses frame representations to transfer the first frame's label to subsequent frames. We term the representation for the jth frame $\mathbf{I}_j$ as $\mathbf{R}_j$. Unlike existing methods that often extract $\mathbf{R}_j$ by a 2D image model, we introduce the 3D UNet backbone from video diffusion models, which includes a temporal axis and processes the entire video sequence as input (see (a)). Our approach improves tracking by integrating temporal motions as shown in (b), where different colors indicate different matching pairs. In (b), the first row shows traditional appearance matching, which relies on visual similarity across frames and may misidentify objects, such as incorrectly matching two deer in the last frame to the same deer in the first frame. In contrast, our work (second row) captures motion patterns among frames, resulting in more accurate tracking. We term our **T**emporal **E**nhanced **D**iffusion tracking method as **TED**. Experiments demonstrate that our TED improves tracking performance across diverse video scenarios, including those with similar-looking objects. (**Best viewed when zoomed in.**)

models, trained to generate consistent videos across frames, capture both the object appearance and the temporal relationships between objects. We show that without additional training, the internal layer outputs of UNet from video diffusion models introduce the temporal reasoning capability that aids tracking in complex situations. For example, as shown in the last column of Figure 1(b), our video diffusion representations successfully track two deer even as they change their positions relative to each other in the video. We term our **T**emporal **E**nhanced **D**iffusion tracking method as **TED**.

Experimental results show that our TED method outperforms 23 popular baseline models, achieving state-of-the-art performance in self-supervised pixel-level object tracking. On the DAVIS dataset for semi-supervised video object segmentation, our TED significantly outperforms SFC (Hu et al., 2022) by 6.4%, SMTC (Qian et al., 2023) by 4.6%, Spa-then-Temp (Li & Liu, 2023) by 3.5%, and DIFT (Tang et al., 2023) by 1.9%. Furthermore, we introduce a challenging task of tracking similar-looking objects, and a new real-world dataset for evaluation, termed YouTube-Similar. Benefiting from the temporal reasoning ability, our TED improves upon DIFT (Tang et al., 2023) by 5.3%. Moreover, our approach achieves state-of-the-art results in human pose tracking. Our work is the first to show that temporal motions learned from video diffusion models can solve perception challenges and significantly improve perception performance. We will release our code and data.

2 Related Work

Learning temporal correspondence is crucial for visual tracking (Tao et al., 2016; Xu & Wang, 2021; Li & Liu, 2023). Due to limited annotations, prior studies have developed methods to learn correspondence in a self-supervised manner (Caron et al., 2021; Qian et al., 2023). Our work contributes to this field of self-supervised correspondence, and we discuss related work as follows.

Temporal correspondence learned from images. Self-supervised learning that trains on image datasets has achieved great success in downstream tasks, including temporal correspondence. Pioneering work, such as MoCo (He et al., 2020) and DINO (Caron et al., 2021), adopt instance discrimination as pretext task which learns similar representations for different augmentations of the same image. DenseCL (Wang et al., 2021) and PixPro (Xie et al., 2021b) further apply contrastive learning to pixel-level, which improve dense prediction tasks. SFC (Hu et al., 2022) boosts performance on temporal correspondence further by fusing image-level and pixel-level representations. Recently, DIFT (Tang et al., 2023) achieves state-of-the-art results in temporal correspondence task by leveraging internal representations from image diffusion models (Rombach et al., 2021). These methods learn intra-frame information and rely on appearance for pixel-level tracking. Our work highlights the limitations of using appearance alone for temporal correspondence and significantly improves tracking by introducing temporal reasoning capabilities.

Temporal correspondence learned from videos. Temporal information in videos provides supervision signals to learn video representations during training. Two widely used pretext tasks for model training are frame reconstruction and cycle-consistency over time. Frame reconstruction tasks involve reconstructing a frame from adjacent frames (Vondrick et al., 2018; Lai & Xie, 2019; Li et al., 2019; Lai et al., 2020), while cycle-consistency tasks track a patch backwards and forward in time to align start and end points (Wang et al., 2019; Jabri et al., 2020). However, these methods often overlook spatial features crucial for creating discriminative and robust representations (Li & Liu, 2023). Recent research integrates spatial with temporal information in model training, such as Spa-then-Temp (Li & Liu, 2023) and SMTC (Qian et al., 2023). Despite incorporating temporal information during model training, our work reveals that existing methods still face challenges in complex scenarios, such as tracking multiple similar-looking objects, as shown in Figure 1. By introducing temporal reasoning ability from video diffusion models to tracking, our approach significantly improves performance across various video scenarios, including those involving similar-looking objects.

Video diffusion models. Diffusion models have significantly advanced image and video generation (Ho et al., 2020; Saharia et al., 2022; Ho et al., 2022; Ruiz et al., 2023). Text-to-image diffusion models (Nichol et al., 2021; Ramesh et al., 2022) allow precise control over generated image content via text prompts, with Stable Diffusion (Rombach et al., 2021) improving generation efficiency and quality by performing diffusion process in latent space. To generate videos with consistent frames, video diffusion models are created by inserting temporal blocks into image diffusion models, which are then trained on video datasets (Blattmann et al., 2023b; Zhang et al., 2023). Representative video diffusion models include Sora (Brooks et al., 2024), ModelScope (Wang et al., 2023), I2VGen-XL (Zhang et al., 2023), and Stable Video Diffusion (Blattmann et al., 2023a). Our work is the first to demonstrate that temporal dynamics learned by video diffusion models can significantly improve tracking performance. Our work highlights the potential of video generative models in tracking tasks beyond their original use in video synthesis.

3 Method

We focus on the video label propagation task and first introduce the background in Section 3.1. We then discuss the challenges faced by previous methods in tracking identical objects in Section 3.2. In Section 3.3, we show how our approach addresses these challenges and improves tracking performance by leveraging temporal context. Our implementation details are provided in Section 3.4.

3.1 Background

Video label propagation task aims to transfer ground truth labels, such as segmentation maps, from the first frame to subsequent frames(Vondrick et al., 2018), as shown in Figure 3. The key is training models to represent frames and establish pixel-level mapping among them (Hu et al., 2022). Due to limited annotations, prior work trains the models in a self-supervised manner with various pretext tasks (Jabri et al., 2020; Li & Liu, 2023). DIFT (Tang et al., 2023) significantly improves tracking performance using latent representations from image diffusion models. We first introduce diffusion models and then discuss how DIFT uses them for tracking.

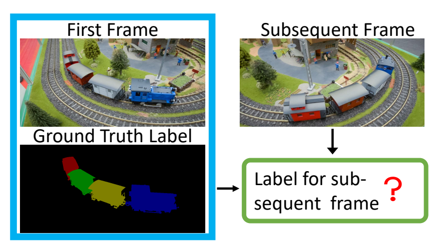


Figure 3: **Video label propagation task** transfers the ground truth label of the first frame to subsequent frames.

Diffusion models have achieved unprecedented success in generating images and videos with rich content (Rombach et al., 2022; Brooks et al., 2024). They are probabilistic models that learn the data distribution $p(\mathbf{x})$ and generate $\mathbf{x}$ from a random Gaussian variable(Nichol et al., 2021), where $\mathbf{x}$ is the image for image diffusion models.

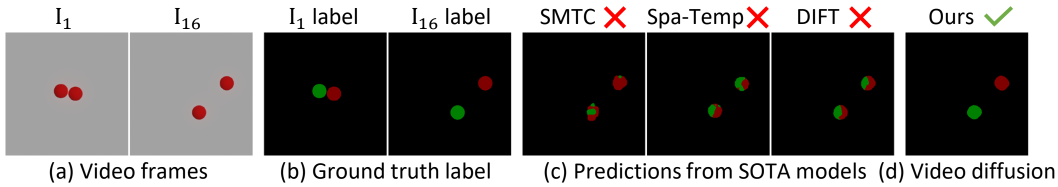


Figure 4: **Video diffusion representations enable tracking objects with identical appearances.** We conduct a controlled study, that we perform object label propagation on videos featuring two independently moving and identical-looking balls, with frames and their ground truth labels shown in (a) and (b). State-of-the-art methods (Qian et al., 2023; Li & Liu, 2023; Tang et al., 2023) fail to distinguish the two balls, leading to incorrect predictions (c). In contrast, our video diffusion representations accurately track both balls despite their identical appearance, as shown in (d).

Diffusion models learn rich visual concepts by recovering signals from corrupted data $\mathbf{x}_\tau$ at varying noise levels (Choi et al., 2022), with loss defined in Eqn. 1:

$$L = \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),\tau}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_\tau, t)\|_2^2\right] \tag{1}$$

where $\boldsymbol{\epsilon}$ is the actual noise corrupting the clean data and $\epsilon_\theta(\mathbf{x}_\tau, t)$ is the noise predicted by the denoising model ϵ_θ. UNet (Ronneberger et al., 2015) is commonly used as the denoising model ϵ_θ.

Noisy $\mathbf{x}_\tau$ is generated by adding noise from a Gaussian distribution $\mathcal{N}(0, 1)$ to the clean data $\mathbf{x}_0$ according to the noise scheduler α_t (Ho et al., 2020), defined as:

$$\mathbf{x}_\tau = \sqrt{\alpha_\tau}\mathbf{x}_0 + \sqrt{1-\alpha_\tau}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1) \tag{2}$$

Here, τ represents the timestep in diffusion process, with larger τ indicating higher noise levels.

Tracking by image diffusion representations. DIFT (Tang et al., 2023) improves video propagation performance using latent representations from image diffusion models (Rombach et al., 2021; Dhariwal & Nichol, 2021). It leverages outputs from internal layers of UNet backbone, defined as:

$$\mathbf{R} = \text{UNet}(\mathbf{x}_\tau, n) \tag{3}$$

where n is the layer index. Specificially, $\mathbf{R}_i = \text{UNet}_i(\mathbf{x}_\tau, n)$, where subscript i indicates image diffusion models. The input $\mathbf{x}_\tau$ is generated using Eqn. 2 at a chosen timestep τ. Since UNet_i processes a single image at a time, DIFT treats video frames as independent images and extracts $\mathbf{R}_i$ for each frame with a single forward pass through the UNet model.

3.2 Challenges for tracking identical objects

Prior studies have achieved impressive results in video label propagation by establishing pixel-level mappings among frames based on frame representations (Jabri et al., 2020; Li & Liu, 2023). For videos with a single object, pixel-level mapping often relies on object appearances, such as the semantic information used in SFC (Hu et al., 2022). However, in videos with multiple similar-looking objects, like tanks with similar fish, establishing accurate correspondence remains challenging and is underexplored in the video label propagation task.

Controlled toy example. We begin with a controlled toy example that tracks two independently moving, identical-looking balls in a video, as shown in Figure 4(a). We use the Kubric simulator (Greff et al., 2022) to create a video dataset with random ball sizes and motions, termed Kubric-Similar. In this dataset, we propagate the segmentation map of each ball from the first frame (see Figure 4(b)) to the subsequent frames. We follow the label propagation procedures in prior studies (Jabri et al., 2020; Li & Liu, 2023), with implementation details in Section 3.4.

We evaluate state-of-the-art models on Kubric-Similar, with results reported in Figure 4(c) and Table 1. Figure 4(c) shows that existing methods struggle with object identity, leading to poor tracking. This aligns with Table 1, where many methods, including DIFT (Tang et al., 2023) that uses image diffusion representations, achieve a J_m around 50%. Note that a J_m around 50% indicates performance no better than random guessing due to the identical size of the two balls. These findings highlight the difficulty of tracking multiple similar-looking objects in temporal correspondence tasks.

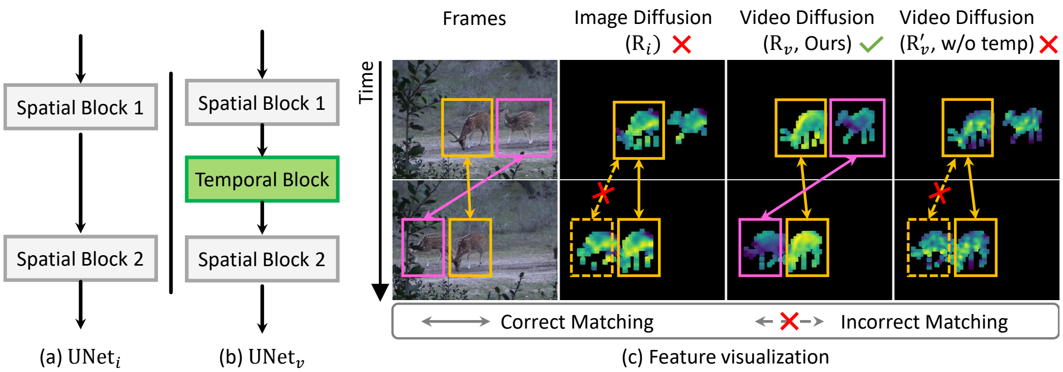


Figure 5: **Track by temporal context.** To understand why video diffusion representation ($\mathbf{R}_v$) excels in tracking similar-looking objects, we compare the UNet backbones of video and image diffusion models. (a) UNet$_i$ from image diffusion models consists of spatial layer blocks that process each image independently. (b) UNet$_v$ is constructed by inserting temporal layer blocks to UNet$_i$ to ensure frame consistency. In (c), we perform principal component analysis (PCA) on the representations from different frames of each model, such as $\tilde{\mathbf{R}}_v^s, \tilde{\mathbf{R}}_v^t = \text{PCA}(\mathbf{R}_v^s \parallel \mathbf{R}_v^t)$ for $\mathbf{R}_v$, where s and t represent different frames. The results reveal that image diffusion representation ($\mathbf{R}_i$) from DIFT (Tang et al., 2023) learns similar features for both deer, leading to incorrect matching. In contrast, our $\mathbf{R}_v$ learns distinguishing features that achieve correct matching. Removing temporal layers from UNet$_v$ results in losing its distinguishing capability, shown in $\mathbf{R}'_v$(w/o temp). By integrating information across frames, $\mathbf{R}_v$ enhances tracking by incorporating temporal context, outperforming $\mathbf{R}_i$ from DIFT (Tang et al., 2023), which is limited to intra-frame information and appearance-based tracking.

Video diffusion representations achieve significant improvement in tracking identical objects. To improve label propagation for identical objects, we replace the image diffusion representations ($\mathbf{R}_i$) in DIFT (Tang et al., 2023) with video diffusion representations ($\mathbf{R}_v$). Specifically, $\mathbf{R}_v$ is obtained by applying UNet$_v$ from video diffusion models following Eqn. 3. Thus, $\mathbf{R}_v = \text{UNet}_v(\mathbf{x}_\tau, n)$, where $\mathbf{x}_\tau$ represents the video sequence of multiple frames. The process of obtaining $\mathbf{R}_v$ is illustrated in Figure 2(a), with additional implementation details in Section 3.4. Figure 4(d) shows that our $\mathbf{R}_v$ accurately tracks both balls, despite their identical appearance, outperforming existing methods. Table 1 further confirms our approach's effectiveness, outperforming DIFT (Tang et al., 2023) by 38.3% in $\mathcal{J}\&\mathcal{F}_\text{m}$.

Table 1: **Results in tracking identical objects.** We perform object label propagation on videos featuring two independently moving, identical-looking balls, as shown in Figure 4(a). Our video diffusion representations achieve state-of-the-art results in tracking identical objects. Colors of the numbers highlight the **best** results.

Model	$\mathcal{J}\&\mathcal{F}_\text{m}(\uparrow)$	$\mathcal{J}_\text{m}(\uparrow)$	$\mathcal{F}_\text{m}(\uparrow)$
MOCO (He et al., 2020)	51.9	61.8	56.8
SimSiam (Chen & He, 2021)	52.8	63.9	58.3
TimeCycle (Wang et al., 2019)	42.6	55.6	49.1
UVC (Li et al., 2019)	58.0	52.6	43.9
CRW (Jabri et al., 2020)	49.5	59.7	54.6
SFC (Hu et al., 2022)	41.8	51.1	46.5
SMTC (Qian et al., 2023)	72.6	68.5	76.7
Spa-then-Temp (Li & Liu, 2023)	44.5	39.8	49.2
DIFT$_{sd}$ (Tang et al., 2023)	46.3	43.4	49.3
DIFT$_{adm}$ (Tang et al., 2023)	52.6	50.3	54.8
Video diffusion ($\mathbf{R}_v$, ours)	**90.9**	**87.2**	**94.5**

3.3 Track objects by temporal context

Motivated by the success of video diffusion representations when tracking identical objects in the toy example above, we will first investigate where the tracking capability comes from. We will then capitalize on the findings and propose a simple and effective method for better tracking.

Where does the ability to track similar-looking objects come from? We hypothesize that this tracking capability stems from temporal context learned during video synthesis. Video diffusion models (Blattmann et al., 2023a) insert temporal layers into image diffusion models to learn temporal dynamics such as motion, ensuring frame consistency in generated videos (see Figure 5 (a)(b)). We denote the UNet representations from video and image diffusion models as $\mathbf{R}_v$ and $\mathbf{R}_i$, respectively.

Table 2: **Results for pixel-level object tracking.** We evaluate our TED method on semi-supervised video object segmentation, and compare it with 24 baseline models, including self-supervised and supervised approaches. Our TED achieves state-of-the-art tracking performance on both the DAVIS and Youtube-Similar datasets, outperforming recent methods by up to 6%. We visualize the tracking results in Figure 6. These results demonstrate the effectiveness of our method in object tracking, even when multiple objects have similar appearances. Colored numbers indicate the **best** results. TED refers to our default setting using $\mathbf{R}_f$, while TED† denotes the setting using $\mathbf{R}_v$.

Supervised	Method	Dataset	DAVIS			Youtube-Similar		
			$\mathcal{J}\&\mathcal{F}_m(\uparrow)$	$\mathcal{J}_m(\uparrow)$	$\mathcal{F}_m(\uparrow)$	$\mathcal{J}\&\mathcal{F}_m(\uparrow)$	$\mathcal{J}_m(\uparrow)$	$\mathcal{F}_m(\uparrow)$
×	InstDis (Wu et al., 2018)	ImageNet w/o labels	66.4	63.9	68.9	-	-	-
×	MoCo (He et al., 2020)	ImageNet w/o labels	65.9	63.4	68.4	48.0	48.5	47.4
×	SimCLR (Chen et al., 2020)	ImageNet w/o labels	66.9	64.4	69.4	37.5	36.9	38.1
×	BYOL (Grill et al., 2020)	ImageNet w/o labels	66.5	64.0	69.0	47.1	47.7	46.5
×	SimSiam (Chen & He, 2021)	ImageNet w/o labels	67.2	64.8	68.8	47.4	47.9	47.0
×	DINO (Caron et al., 2021)	ImageNet w/o labels	71.4	67.9	74.9	63.9	63.0	64.7
×	DetCo (Xie et al., 2021a)	ImageNet w/o labels	65.7	63.3	68.1	41.4	42.0	40.9
×	DenseCL (Wang et al., 2021)	ImageNet w/o labels	61.4	60.0	62.9	46.4	46.5	46.3
×	PixPro (Xie et al., 2021b)	ImageNet w/o labels	57.5	56.6	58.3	45.6	45.9	45.3
×	DIFT$_{adm}$ (Tang et al., 2023)	ImageNet w/o labels	75.7	72.7	78.6	60.7	59.8	61.7
×	DIFT$_{sd}$ (Tang et al., 2023)	LAION	70.0	67.4	72.5	56.3	55.8	56.7
×	Colorization (Vondrick et al., 2018)	Kinetic	34.0	34.6	32.7	-	-	-
×	VINCE (Gordon et al., 2020)	Kinetic	65.2	62.5	67.8	44.9	45.4	44.3
×	VFS (Xu & Wang, 2021)	Kinetic	68.9	66.5	71.3	57.3	57.1	57.5
×	UVC (Li et al., 2019)	Kinetic	60.9	59.3	62.7	49.7	49.8	49.7
×	CRW (Jabri et al., 2020)	Kinetic	67.6	64.8	70.2	52.0	52.3	51.6
×	CorrFlow (Lai & Xie, 2019)	OxUvA	50.3	48.4	52.2	39.6	40.0	39.3
×	TimeCycle (Wang et al., 2019)	VLOG	48.7	46.4	50.0	39.8	41.3	38.2
×	MAST (Lai et al., 2020)	YT-VOS	65.5	63.3	67.6	-	-	-
×	SMTC (Qian et al., 2023)	YT-VOS	73.0	69.4	76.6	57.5	57.2	57.9
×	SFC (Hu et al., 2022)	ImageNet, YT-VOS	71.2	68.3	74.0	55.5	55.3	55.7
×	Spa-then-Temp (Li & Liu, 2023)	ImageNet, YT-VOS	74.1	71.1	77.1	59.6	59.2	60.1
×	**TED† (Ours, R_v)**	Web-Vid	66.3	63.4	69.1	62.0	61.5	62.5
×	**TED (Ours, R_f)**	ImageNet, Web-Vid	**77.6**	**74.4**	**80.8**	**66.0**	**65.1**	**67.0**
✓	OSVOS (Caelles et al., 2017)	ImageNet, DAVIS	60.3	56.6	63.9	-	-	-
✓	OnAVOS (Voigtlaender & Leibe, 2017)	ImageNet, DAVIS	65.4	61.6	69.1	-	-	-
✓	CFBI+ (Yang et al., 2020)	YT-VOS, DAVIS	82.8	80.1	85.5	-	-	-

We examine the properties of $\mathbf{R}_v$ and $\mathbf{R}_i$ using principal component analysis (PCA), as shown in Figure 5(c), where two moving deer change their relative positions over time. We will show that $\mathbf{R}_v$ learns distinguishing features for two deer even if they have similar appearances, while $\mathbf{R}_i$ learns similar features for both deer.

We perform PCA on pairs of frames for each model, such as $\tilde{\mathbf{R}}_v^s, \tilde{\mathbf{R}}_v^t = \text{PCA}(\mathbf{R}_v^s \parallel \mathbf{R}_v^t)$ for $\mathbf{R}_v$ where s and t represent different frames. Figure 5(c) shows that $\mathbf{R}_i$ of DIFT (Tang et al., 2023) learn similar features for both deer, leading to incorrect matching. In contrast, $\mathbf{R}_v$ learns distinguishing features for the two deer that enable correct matching. We then remove the temporal blocks from UNet$_v$ and recomputed $\mathbf{R}_v$, termed $\mathbf{R}_v'$. Interestingly, Figure 5(c) shows that $\mathbf{R}_v'$ loses the distinguishing features between the two deer. Unlike $\mathbf{R}_i$ which only uses intra-frame information, the temporal layers in UNet$_v$ (like temporal attention layers) enable $\mathbf{R}_v$ to integrate information across multiple video frames, introducing temporal motion to tracking. We compare our temporal motion matching using $\mathbf{R}_v$ to the appearance matching of $\mathbf{R}_i$ in Figure 2(b). Our results and discussions demonstrate the superiority of $\mathbf{R}_v$ in using temporal context for tracking.

Using $\mathbf{R}_v$ for better tracking. Our investigations show that video diffusion representations ($\mathbf{R}_v$) capture temporal context, crucial for tracking identical objects. Since temporal context is orthogonal to appearance information, it complements prior tracking methods like image diffusion ($\mathbf{R}_i$). As shown in Eqn. 4, we employ a simple concatenation of the representations from video and image diffusion models in later experiments:

$$\mathbf{R}_f = \texttt{Concat}\,(\alpha\|\mathbf{R}_v\|_2,\ (1-\alpha)\|\mathbf{R}_i\|_2) \tag{4}$$

where $\|\cdot\|$ denotes L2 normalization and α is a hyperparameter between 0 and 1. We term our **T**emporal **E**nhanced **D**iffusion tracking method as **TED**. We use $\mathbf{R}_f$ by default and denote the setting that uses $\mathbf{R}_v$ as TED† for distinction. We will show our TED achieves state-of-the-art tracking results.

3.4 Implementation details

Video label propagation. Our work follows prior studies (Jabri et al., 2020; Caron et al., 2021; Tang et al., 2023) for the evaluation protocol of label propagation, which includes representation extraction and label prediction stage, as shown in Algorithm 1. We first obtain frame representations R_f using video and image diffusion models. To predict the label of current frame, similar pixel pairs between current and previous frames are identified by computing the similarities of their representations. Each pixel in the current frame is then labeled by aggregating the labels of similar pixels from previous frames, weighted by their pixel similarity. More experimental setups are detailed in Appendix B.1.

Appearance representations. Following (Tang et al., 2023), we use the output from the internal layers of UNet_i as the appearance representation $\mathbf{R}_i$, following $\mathbf{R}_i = \text{UNet}_i(\mathbf{x}_\tau, n)$. $\mathbf{x}_\tau$ represents each video frame, generated according to Eqn. 2 with an empirically determined τ. We process each frame through a single forward pass of UNet_i. Our framework accommodates any pre-trained image diffusion model for $\mathbf{R}_i$, using ADM (Dhariwal & Nichol, 2021) by default. We also investigate other models such as Stable Diffusion (Rombach et al., 2022).

Temporal representations. We obtain $\mathbf{R}_v$ following $\mathbf{R}_v = \text{UNet}_v(\mathbf{x}_\tau, n)$ as shown in Fig 2(a). The key difference in obtaining $\mathbf{R}_v$ compared to $\mathbf{R}_i$ is using UNet_v from video diffusion models, which process video sequence of multiple frames as $\mathbf{x}_\tau$. Since current video diffusion models accept limited frames as input, long videos are split into subsequences. $\mathbf{R}_v$ is then obtained for each subsequence through a one-pass forward process in UNet_v. Our framework supports any off-the-shelf pre-trained video diffusion model for $\mathbf{R}_v$, using I2VGen-XL (Zhang et al., 2023) by default. We also explore additional models like Stable Video Diffusion (Blattmann et al., 2023a).

4 Experiments

4.1 Experimental setup

We evaluate our TED method on the video label propagation task, and compare it with 24 baseline models. Our work uses video representations from pre-trained diffusion models, and does not require additional training.

- **Pretrained self-supervised learning models.** We evaluate 9 self-supervised models pre-trained on ImageNet known for strong temporal correspondence performance: 6 instance discrimination models like MoCo (He et al., 2020) and 3 dense contrastive learning models, such as DenseCL (Wang et al., 2021).
- **Image diffusion model representations.** We compare with DIFT (Tang et al., 2023), which leverages representations from image diffusion models for temporal correspondence.
- **Task-specific models (self-supervised).** We include 11 self-supervised models tailored for temporal correspondence tasks, trained by pretext tasks like frame reconstruction (e.g., UVC (Li et al., 2019)), cycle consistency (e.g., CRW (Jabri et al., 2020)), and video contrastive learning (e.g., VFS (Xu & Wang, 2021)). We also include recent methods such as SMTC (Qian et al., 2023) and Spa-then-Temp (Li & Liu, 2023).
- **Task-specific models (supervised).** We compare our method with 3 supervised approaches that utilize labeled data during training, such as CFBI+ (Yang et al., 2020).

Evaluation datasets. We evaluate TED on the semi-supervised video object segmentation task, which propagates the object segmentation from the first frame to subsequent frames. We evaluate widely-used DAVIS-2017 (Pont-Tuset et al., 2017) which includes 30 videos from various scenarios. To test the tracking ability for similar-looking objects, we introduce the Youtube-Similar dataset, composed of 28 videos from Youtube-VOS (Xu et al., 2018) that feature multiple similar-looking objects. Following (Tang et al., 2023), we report region-based similarity (J_m) and contour-based accuracy (F_m). More dataset details are provided in the Appendix B.3.

4.2 Experimental results

Quantitative results. We compared our TED to 24 baseline models on the DAVIS and Youtube-Similar dataset, with results detailed in Table 2. Our TED achieves the **state-of-the-art** tracking

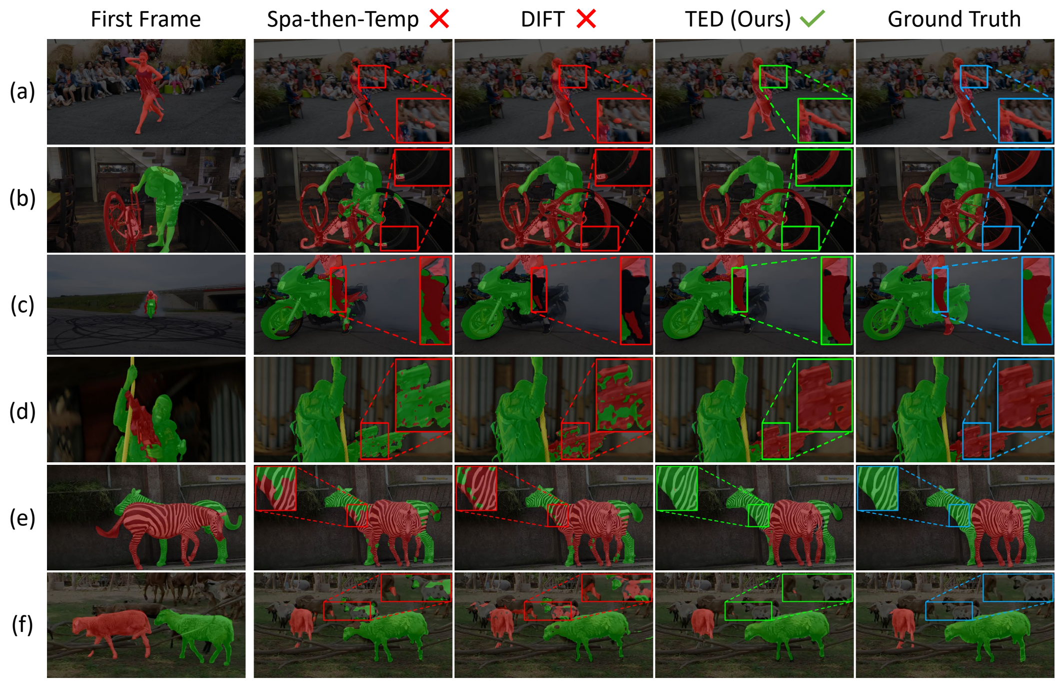


Figure 6: **Predictions for pixel-level object tracking.** We evaluate our TED method on semi-supervised video object segmentation, which propagates object segmentation maps from the first frame to subsequent frames. Our TED consistently outperforms state-of-the-art methods (Li & Liu, 2023; Tang et al., 2023) on the DAVIS (Figure a-d) and YouTube-Similar (Figure e-f) datasets, aligning with the results in Table 2. Notably, our TED delivers more accurate predictions in scenarios with complex deformations (a) and viewpoint changes (b), while Spa-then-Temp (Li & Liu, 2023) and DIFT (Tang et al., 2023) struggle with tracking completeness, such as the missing arm in (a). Our TED also excels in multi-object scenarios, delivering superior tracking for interacting objects (c-d) and similar-looking objects (e-f). In contrast, Spa-then-Temp (Li & Liu, 2023) and DIFT (Tang et al., 2023) have mislabeling issues, such as incorrect labels for the gun in (d) and misaligned labels for sheep in the background (f). These results show that our TED significantly improves tracking performance, highlighting the benefits of incorporating temporal reasoning into tracking. (**Best viewed when zoomed in.**)

performance on both datasets, surpassing recent methods by up to **6%**. Specifically, on the DAVIS dataset, our method outperforms SFC (Caron et al., 2021) by 6.4%, SMTC (Qian et al., 2023) by 4.6%, Spa-then-Temp (Li & Liu, 2023) by 3.5%, and DIFT (Tang et al., 2023) by 1.9%. On the Youtube-Similar dataset, our TED shows an even greater improvement, exceeding Spa-then-Temp (Li & Liu, 2023) by 6.4% and DIFT (Tang et al., 2023) by 5.3%. These improvements highlight the effectiveness of our method in object tracking, even for challenging settings with multiple similar-looking objects.

Visualizations. We present our tracking results alongside those from state-of-the-art methods in Figure 6, with results for DAVIS shown in Figure 6(a-d) and for YouTube-Similar in Figure 6(e-f). Our TED outperforms existing studies on both datasets, aligning with Table 2. Our TED effectively handles complex deformations (a) and viewpoint changes (b), outperforming Spa-then-Temp (Li & Liu, 2023) and DIFT (Tang et al., 2023), which struggle with tracking elements like the human arm in Figure 6(a). Additionally, our TED excels in multiple-object scenarios, such as interacting objects (c-d) and similar-looking objects (e-f), whereas Spa-then-Temp and DIFT often confuse different objects, leading to incorrect label assignments. For instance, in Figure 6(d), Spa-then-Temp (Li & Liu, 2023) incorrectly labels the gun as a human, and DIFT (Tang et al., 2023) shows significant errors in the predicted contour. In Figure 6 (f), featuring multiple sheep, both Spa-then-Temp (Li & Liu, 2023) and DIFT (Tang et al., 2023) mistakenly align the object label to the sheep in the background. Our TED consistently achieves more accurate tracking results across these scenarios, demonstrating significant performance improvements through enhanced temporal reasoning.

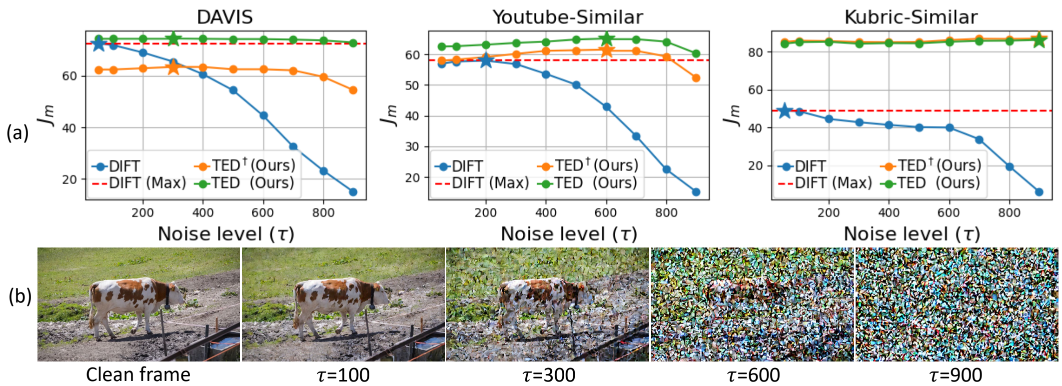


Figure 7: **How and why does t influence tracking.** We present tracking results in (a) using diffusion representations obtained at varying noise levels τ (see Eqn. 2 and Eqn. 3), with higher τ indicating more noise (b). TED uses combined $\mathbf{R}_f$ defined in Eqn. 4, and TED† uses video diffusion representation $\mathbf{R}_v$. The best result for each method is marked with a star and the best result for DIFT (Tang et al., 2023) across all τ is indicated as a red dashed line. Using image diffusion representations ($\mathbf{R}_i$), DIFT peaks at low noise ($\tau \leq 200$) and deteriorates as noise increases. This is due to its reliance on appearance for tracking, which becomes almost unavailable at high noises. In contrast, TED† (using $\mathbf{R}_v$) excels at higher τ values, peaking at τ=600 on Youtube-Similar and τ=900 on Kubric-Similar where the input video is heavily corrupted (b). The high accuracy at high noise levels is because $\mathbf{R}_v$ learns coarse-grained motions that enable tracking similar-looking objects, such as object positions. When the video input is less noisy, the diffusion model is trained to denoise appearance details, where motion feature may not be so prioritized, leading to performance decrease at low noise levels. Our TED consistently outperforms DIFT (Tang et al., 2023) across various τ values on all datasets, demonstrating the superiority of incorporating temporal information into tracking.

4.3 Ablation studies and discussions

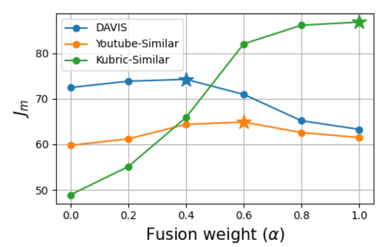


Figure 8: **Fusion weight α.** Our TED outperforms DIFT (Tang et al., 2023)(α=0.0) on all datasets.

How and why does τ influence tracking. We obtain frame representations from diffusion models as defined in Eqn. 3, with the UNet input $\mathbf{x}_\tau$ generated according to Eqn. 2. Following DIFT (Tang et al., 2023), we empirically determine the noise level t to produce $\mathbf{x}_\tau$. We investigate the impact of noise level τ on tracking performance in Figure 7(a), where a higher τ indicates more noise (Figure 7(b)). In Figure 7(a), Kubric-Similar is a dataset featuring independently moving and identical-looking balls, defined in Section 3.2. We mark the best result for each method with a star. TED uses combined $\mathbf{R}_f$ defined in Eqn. 4, and TED† uses video diffusion representation $\mathbf{R}_v$. Using image diffusion representations ($\mathbf{R}_i$), DIFT (Tang et al., 2023) achieves the best result at low noise ($\tau \leq 200$) and decreases rapidly as noise increases due to diminishing availability of appearance information. In contrast, our TED† with $\mathbf{R}_v$ peaks at a higher t and maintains robust tracking over a much broader range of τ. Notably, TED† reaches its best performance at τ=600 on Youtube-Similar and τ=900 on Kubric-Similar, where the input video is heavily corrupted and appearance information is almost unavailable as shown in Figure 7(b). These results suggest that $\mathbf{R}_v$ encodes temporal motion that can be used for tracking at higher noise levels. Moreover, our TED with $\mathbf{R}_f$ consistently outperforms DIFT (Tang et al., 2023) across a wide τ range, demonstrating the effectiveness of our TED by integrating temporal dynamics into tracking.

Diffusion models solve different tasks at different noise levels during training (Choi et al., 2022). When the video input is corrupted at high noise levels, video diffusion models are trained to solve the hard task that learning coarse-grained signals in the video, such as motion (like the change of object positions among frames). Therefore, its representation encodes rich motion information that enables tracking similar-looking objects. When the video input is less noisy, the diffusion model is trained to denoise appearance details, where motion features may not be so prioritized, leading to performance decrease at low noise levels.

Fusion weight α. To utilize both temporal motion and appearance for better tracking, our TED combines $\mathbf{R}_v$ and $\mathbf{R}_i$ into $\mathbf{R}_f$ as defined in Eqn. 4. Figure 8 shows the tracking results with varying fusion weight α, where a higher α increases the contribution of $\mathbf{R}_v$. $\mathbf{R}_f$ reduces to $\mathbf{R}_i$ when α=0.0 and to $\mathbf{R}_v$ when α=1.0. We mark the best result on each dataset with a star in Figure 8. Our TED achieves the best results with medium α values of 0.4 for DAVIS and 0.6 for Youtube-Similar, demonstrating that the integration of appearance and temporal information improves tracking performance. For Kubric-Similar, TED performs best with α=1.0, reflecting the dataset's unique characteristics of containing identical objects where appearance information from $\mathbf{R}_i$ does not provide additional value for tracking. On all datasets, our TED consistently outperforms DIFT (Tang et al., 2023) (α=0.0), highlighting the advantage of our work by introducing temporal motions to tracking.

Feature layers for video diffusion representations. We use $\mathbf{R}_v$ from internal layers of the UNet in video diffusion models for video label propagation, as illustrated in Figure 2 and Eqn. 3. Following (Tang et al., 2023), we use the decoder representations from UNet and report the tracking results of TED† on DAVIS using $\mathbf{R}_v$ from different decoder blocks in Table 3. Table 3 shows that the medium block (block 2) yields the best performance among all blocks.

Table 3: **Ablation study on UNet blocks.** TED† achieves the best tracking results using $\mathbf{R}_v$ from block 2.

Block	J_m&F_m	J_m	F_m
0	24.8	28.2	21.4
1	47.6	52.7	42.5
2	66.3	63.4	69.1
3	31.5	27.2	35.8

Different diffusion models. We evaluate the tracking results of TED using $\mathbf{R}_f$ obtained from different video and image diffusion models on the DAVIS dataset, as shown in Table 4. We investigate video diffusion models like Stable Video Diffusion (SVD) (Blattmann et al., 2023a) and I2VGen-XL (I2V) (Zhang et al., 2023), image diffusion models like Stable Diffusion (SD) (Rombach et al., 2022) and ADM (Dhariwal & Nichol, 2021). Our TED achieves the best tracking performance when using video diffusion representations from I2VGen-XL (Zhang et al., 2023) and image diffusion representations from ADM (Dhariwal & Nichol, 2021), which is used as the default setting in the paper.

Table 4: **Pretrained diffusion models for TED**. Our TED achieves the best tracking results using representations from I2VGen-XL and ADM.

Video	Image	$\mathcal{J}$&$\mathcal{F}_\mathrm{m}(\uparrow)$	$\mathcal{J}_\mathrm{m}(\uparrow)$	$\mathcal{F}_\mathrm{m}(\uparrow)$
SVD	SD	71.5	68.9	74.1
SVD	ADM	76.6	73.6	79.7
I2V	SD	71.7	69.0	74.5
I2V	ADM	**77.6**	**74.4**	**80.8**

Results on human pose tracking. In addition to video object segmentation, we test our method on the JHMDB benchmark (Jhuang et al., 2013), which tracks 15 human pose keypoints in 268 videos. We follow the evaluation protocol of prior studies (Li et al., 2019; Jabri et al., 2020; Li & Liu, 2023), and report the percentage of correctly tracked keypoints (PCK) for JHMDB. We compare our method with baseline models in Table 5. Table 5 shows that our approach achieves state-of-the-art performance in the human pose tracking task.

Table 5: **Results on JHMDB dataset.** Our method achieves state-of-the-art performance in human pose tracking.

Method	PCK@0.1	PCK@0.2
SFC (Hu et al., 2022)	61.9	83.0
SMTC (Qian et al., 2023)	62.5	84.1
DIFT (Tang et al., 2023)	63.4	84.3
Spa-then-Temp (Li & Liu, 2023)	66.4	84.4
TED (Ours)	**68.3**	**85.8**

5 Conclusion

In this work, we leverage latent representations from video diffusion models for pixel-level tracking. Benefiting from video diffusion models' ability to incorporate information across multiple frames, our work introduces temporal reasoning to the tracking tasks. Without additional training, our method improves tracking performance in various video scenarios, even enabling tracking of similar-looking objects where previous methods struggle. Experimental results show that our approach achieves state-of-the-art tracking performance, outperforming recent studies by up to 6 points. Our work highlights the potential of video generative models in tracking applications beyond their original use in video synthesis task.

REFERENCES

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

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*, 2023a.

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*, pp. 22563–22575, 2023b.

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. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 221–230, 2017.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11472–11481, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Tobias Gerstenberg, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. How, whether, why: Causal judgments as counterfactual contrasts. In *CogSci*, 2015.

Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Yingdong Hu, Renhao Wang, Kaifeng Zhang, and Yang Gao. Semantic-aware fine-grained correspondence. In *European Conference on Computer Vision*, pp. 97–115. Springer, 2022.

HuggingFace. Diffusers documentation. `https://huggingface.co/docs/diffusers/index`, 2024.

Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020.

Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 3192–3199, 2013.

Zihang Lai and Weidi Xie. Self-supervised learning for video correspondence flow. In *BMVC*, 2019.

Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2020.

Rui Li and Dong Liu. Spatial-then-temporal self-supervised learning for video correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2279–2288, 2023.

Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. *Advances in Neural Information Processing Systems*, 32, 2019.

Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16675–16687, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

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*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Mohammadreza Salehi, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. Time does tell: Self-supervised time-tuning of dense image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16536–16547, 2023.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36: 1363–1389, 2023.

Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1420–1429, 2016.

Tomer David Ullman. *On the nature and origin of intuitive theories: learning, physics and psychology*. PhD thesis, Massachusetts Institute of Technology, 2015.

Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017.

Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 391–408, 2018.

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.

Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2566–2576, 2019.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3024–3033, 2021.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8392–8401, 2021a.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16684–16693, 2021b.

Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10075–10085, 2021.

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

Zongxin Yang, Yuhang Ding, Yunchao Wei, and Yi Yang. Cfbi+: Collaborative video object segmentation by multi-scale foreground-background integration. *The 2020 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2020.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

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. *Advances in Neural Information Processing Systems*, 36, 2024.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qing, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. 2023.

A Discussions

A.1 Advantages of our work over DIFT

We clarify and highlight the advantages of our work over the state-of-the-art DIFT (Tang et al., 2023) as follows.

- **We solve a task that tracks similar-looking objects which DIFT cannot solve.** Tracking similar-looking objects in label propagation is a very fundamental task in the field. Since DIFT learns only appearance features, it fails to track similar-looking objects.
- **Our work uses temporal motions learned from video diffusion models in tracking, providing new insights into how motion-based tracking emerges.** Our experiments and analysis show that temporal layers in video diffusion models enable motion-aware features necessary for tracking similar-looking objects, which are absent in DIFT.
- **Improved tracking accuracy across various scenarios.** Our work significantly outperforms DIFT in tracking performance in various videos, such as those with severe object deformation, achieving 1.9% higher accuracy on DAVIS and 5.3% on YouTube-Similar.

A.2 Advantage of our work in learning temporal features

We clarify and highlight the advantages of our work over prior studies in learning temporal features as follows.

- **Better representations obtained by solving a harder generative task.** Previous methods are trained on easier tasks that always have shortcuts. For example, mismatched patches in Wang et al. (2019); Jabri et al. (2020) or objects in Gordon et al. (2020); Xu & Wang (2021) with similar appearances can also yield low training loss. In contrast, our video diffusion models are trained to fully reconstruct every pixel from noisy inputs, enabling better representation learning.
- **Advanced temporal attention *vs.* simple pairwise correlation.** During training, prior methods learn temporal features by simple correlations between spatial features across frames (Vondrick et al., 2018; Wang et al., 2019; Li et al., 2019; Lai & Xie, 2019; Lai et al., 2020; Qian et al., 2023; Li & Liu, 2023), which fail to distinguish similar-looking objects. In contrast, our video diffusion model uses temporal attention layers to integrate multiple frames, enabling advanced reasoning in complex scenarios like the deer with changing positions in Figure 1.
- **Significantly improved tracking accuracy.** Quantitative results and visualization show that our method significantly improves the tracking performance compared to prior studies, by more than 3.5% on DAVIS and 6.4% on YouTube-Similar.

B Experimental setups

B.1 Video label propagation

In this work, we evaluate the video label propagation task, which predicts pixel-level labels for subsequent video frames given the ground-truth label of the first frame. We follow the evaluation protocol of prior studies (Jabri et al., 2020; Caron et al., 2021; Tang et al., 2023), as detailed in Algorithm 1. Pixel-level labels for each frame are predicted based on the frame representations and labels of previous frames. For the current frame, we first identify similar pixel pairs between this frame and the previous frames by computing the similarities of their pixel representations. Labels for the current frame are then predicted by aggregating the labels of similar pixels from previous frames, weighted by their pixel similarity. A key advantage of TED over prior studies is that it generates frame representations by inputting the video sequence into the 3D UNet$_v$, which encodes the temporal motions learned in video generation and significantly improves tracking accuracy. To handle videos longer than the sequence length limit of UNet$_v$, we split each video into multiple sequences and process each sequence separately. An optional technique to improve accuracy is allowing overlapping frames among sequences. Another optional technique is using a batch of random noise to obtain an averaged representation for each video following DIFT (Tang et al., 2023).

Algorithm 1: Temporal Enhanced Diffusion tracking (TED)

Input: Video frames I_1 to I_N; Ground-truth label for the first frame L_1; Video diffusion model with UNet$_v$; Image diffusion model with UNet$_i$.

Output: Label predictions L_2 to L_N for frames I_2 to I_N.

Let d be the sequence length defined by UNet$_v$, split all video frames to $n = \lfloor \frac{N}{d} \rfloor$ sequences;
Initialize queue $Q = \emptyset$ to store the representations and labels of the previous p frames;
for *each sequence* $j = 0$ **to** $n - 1$ **do**
  Select the frames $I_{1+j \cdot d}$ to $I_{(j+1) \cdot d}$ as the current sequence;
  Step 1: Computation of Frame Representations
  Compute the video diffusion representation R_v using a single forward pass of UNet$_v$:
    $R_{v,1+j\cdot d}, \ldots, R_{v,(j+1)\cdot d} = \text{UNet}_v(I_{1+j\cdot d}, \ldots, I_{(j+1)\cdot d})$;
  Compute the image diffusion representation R_i using d forward passes of UNet$_i$:
    for *each frame* I_k *in* $I_{1+j\cdot d}$ *to* $I_{(j+1)\cdot d}$ **do**
      $R_{i,k} = \text{UNet}_i(I_k)$;
  Compute the fused representation R_f following Eqn. 4:
    $R_f = \texttt{Concat}\,(\alpha \|R_v\|_2,\ (1-\alpha)\|R_i\|_2)$;
  Step 2: Label Prediction
  if $j = 0$ **then**
    Add $(R_{f,1}, L_1)$ of the first frame to the queue Q;
  for *each frame* I_k *in the sequence from* $I_{1+j\cdot d}$ *to* $I_{(j+1)\cdot d}$ **do**
    if $k = 1$ **then**
      Skip the first frame since the ground truth label is already provided ;
    Compute the pixel similarity matrix A between pixel representations of current frame $R_{f,k}$ and previous frames $R \in Q$;
    for *each pixel in the frame* I_k **do**
      Retain only the similarities for spatially neighboring pixels in A;
      Apply top-κ filtering to retain the strongest similarities and set the remaining values in A to zero;
    Predict the labels for the current frame k by propagating the labels from the most similar pixels in previous frames, weighted by their pixel similarity:
      $L_k = A\cdot$ (labels $L \in Q$);
    Add $(R_{f,k}, L_k)$ to the queue Q;
    if *the size of* Q *exceeds the maximum allowed size* p **then**
      Remove the oldest entry from the queue Q;
return L_2 *to* L_N.

B.2 PRETRAINED DIFFUSION MODELS

In our work, we utilize pretrained diffusion models without additional training. Our framework supports any pretrained diffusion model and we use open-sourced checkpoints for our experiments. For video diffusion models, we use the official weights of Stable Video Diffusion (Blattmann et al., 2023a) and I2VGen-XL (Zhang et al., 2023) available on Hugging Face (HuggingFace, 2024). For image diffusion models, we use pretrained weights from Hugging Face for Stable Diffusion (Rombach et al., 2022) (version 2-1) and from the official GitHub repository for ADM (Dhariwal & Nichol, 2021). We follow the configurations of DIFT (Tang et al., 2023) and summarize as in Table 6.

Table 6: Experimental setups of TED in video label propagation task.

Dataset	Video diffusion			Image diffusion			Fusion	Softmax	Propagation	k for
	Model	Timestep	Block	Model	Timestep	Block	weight	temp	radius	top-k
DAVIS	I2VGen-XL	300	2	ADM	51	7	0.4	0.2	15	10
Youtube-Similar	I2VGen-XL	600	2	ADM	51	7	0.6	0.1	15	10

B.3 Evaluation datasets

We evaluate TED on the semi-supervised video object segmentation task using three datasets: DAVIS-2017 (Pont-Tuset et al., 2017), Youtube-Similar, and Kubric-Similar. Figure 9 shows video examples from each dataset.

- **DAVIS-2017** (Pont-Tuset et al., 2017): A widely used benchmark for semi-supervised object segmentation. Following prior work (Caron et al., 2021; Tang et al., 2023), we use the val subset, which includes 30 videos with 2023 frames and 59 annotated objects.
- **Youtube-Similar**: We propose this benchmark to evaluate tracking on multiple similar-looking objects. It includes 28 videos from Youtube-VOS (Xu et al., 2018) with 839 frames and 69 annotated objects.
- **Kubric-Similar**: We use Kubric simulator (Greff et al., 2022) to generate this dataset for tracking identical objects. Each of the 14 videos contains two identical balls with random sizes and movements, totaling 224 frames and 28 objects.

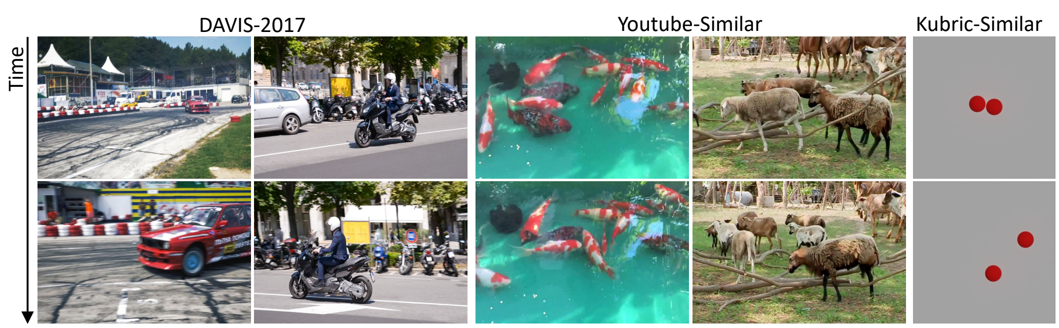


Figure 9: **Dataset examples.** We present video examples from various evaluation datasets. Following prior work (Caron et al., 2021; Tang et al., 2023), we evaluate our method on the widely-used DAVIS-2017 dataset (Pont-Tuset et al., 2017), shown in the first two columns of the figure. For the first time, we propose the challenging task of tracking multiple similar-looking objects in video label propagation. To assess model performance in this setting, we introduce two new datasets: Youtube-Similar (the third and fourth columns) and Kubric-Similar (the fifth column).

B.4 Feature visualization

In Section 3.3, we use PCA (Maćkiewicz & Ratajczak, 1993) to reduce the dimension of pixel representations for visualization. Figure 5(c) visualizes the representations after PCA, where similar colors indicate similar pixel representations. If different objects have distinct pixel colors, it indicates they are successfully distinguished from each other. Figure 5(c) shows that our work succeeds in distinguishing and tracking similar-looking objects (third column), unlike DIFT which learns similar pixel representations for different objects and fails in tracking (second column). These results highlight the effectiveness of temporal motions in our work for tracking, which DIFT lacks.

C Additional results

C.1 Computation cost anlysis

We compare computation cost of our method with DIFT (Tang et al., 2023) in Table 7. We track a 100-frame video, reporting average time per frame and maximum GPU memory. Our TED (efficient) outperforms DIFT (best) by 1.5% in accuracy with similar speed and slightly higher memory use, while TED (best) achieves higher accuracy at greater computation cost. In real applications, users can choose the version based on their requirements on accuracy and efficiency.

We introduce the setups for computation cost analysis as follows. We test the model on a single NVIDIA TITAN RTX GPU using a 100-frame DAVIS video. Following DIFT, we introduce two TED versions, efficient and best, based on whether to use the optional techniques. For DIFT, the optional technique involves averaging representations using a batch of noise. For TED, it includes both averaged representations and overlapping frames among sequences, as discussed in Appendix B.1.

Our work demonstrates, for the first time, that motions learned from video diffusion models can solve perception challenges and achieve state-of-the-art results. Our work offers new insights for diffusion and tracking, benefiting both communities. We believe our method can be further accelerated with future research on diffusion model acceleration as well as advances in computing and resources.

Table 7: **Computation cost analysis.** Our TED (efficient) outperforms DIFT (best) by 1.5% in accuracy with similar speed and slightly higher memory use, while TED (best) achieves higher accuracy at greater computation cost. Here, the time refers to the duration required to track a single image.

Model	Version	Optional Techniques	Accuracy	Time (s)	FPS	Memory (GB)
DIFT (Tang et al., 2023)	Efficient	No	74.7	0.73	1.37	5.53
	Best	Yes	75.7	1.36	0.74	9.25
TED(ours)	Efficient	No	77.2	1.21	0.82	11.65
	Best	Yes	77.6	2.24	0.47	15.20

C.2 Discussions on the training dataset

To investigate the influence of training dataset on the tracking results, we train image diffusion model from DIFT (Tang et al., 2023) on the same training dataset as our video diffusion model for comparison. Table 8 shows that without temporal modeling, training on additional video data fails to track similar-looking objects, indicated by a low $\mathcal{J}\&\mathcal{F}_\mathrm{m}$ of 43.8% on Kubric-Similar. Web-Vid (Bain et al., 2021) has lower individual image quality (Guo et al., 2024), such as motion blur and watermarks. Fine-tuning DIFT on Web-Vid even reduces performance. In contrast, our TED achieves significant improvements using video diffusion models and effectively distinguishes similar-looking objects, demonstrating the importance of learning temporal motions from video diffusion models for tracking.

Table 8: **Fine-tune DIFT's image diffusion models on video datasets**. DIFT fails to track similar-looking objects even when trained on the same datasets as our video diffusion models. This is because image diffusion models learn only appearance features from video datasets, lacking the temporal motion information critical for tracking.

Version	Model	Dataset	Kubric-Similar			Youtube-Similar			DAVIS		
			$\mathcal{J}\&\mathcal{F}_\mathrm{m}(\uparrow)$	$\mathcal{J}_\mathrm{m}(\uparrow)$	$\mathcal{F}_\mathrm{m}(\uparrow)$	$\mathcal{J}\&\mathcal{F}_\mathrm{m}(\uparrow)$	$\mathcal{J}_\mathrm{m}(\uparrow)$	$\mathcal{F}_\mathrm{m}(\uparrow)$	$\mathcal{J}\&\mathcal{F}_\mathrm{m}(\uparrow)$	$\mathcal{J}_\mathrm{m}(\uparrow)$	$\mathcal{F}_\mathrm{m}(\uparrow)$
Original	DIFT	ImageNet	52.6	50.3	54.8	60.7	59.8	61.7	75.7	72.7	78.6
Finetune on Web-Vid	DIFT	ImageNet, Web-Vid	43.8	40.1	47.4	58.9	58.3	59.4	72.9	70.1	75.6
Ours	TED	ImageNet, Web-Vid	**90.4**	**86.9**	**94.0**	**66.0**	**65.1**	**67.0**	**77.6**	**74.4**	**80.8**

C.3 RESULTS OF TIME-TUNING METHOD

We use Time-Tuning features (Salehi et al., 2023) for video label propagation task and find that it fails to distinguish similar-looking objects in our work, as shown in Figure 10. This failure is because Time-Tuning is trained to learn semantic features for semantic segmentation task, as shown in Figure 3 of the original paper (Salehi et al., 2023), which lacks object motions needed in tracking similar-looking objects.

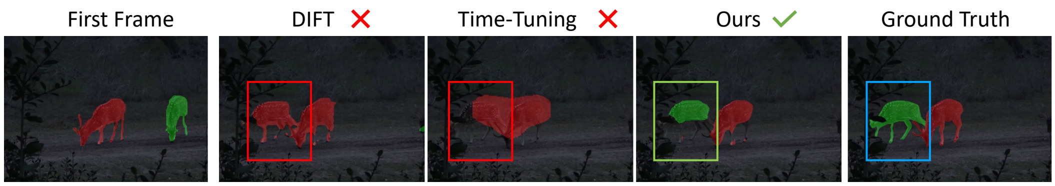


Figure 10: Time-Tuning fails to distinguish multiple similar-looking objects.

C.4 RESULTS WITH ADDITIONAL DINO FEATURES

Prior work (Zhang et al., 2024) shows that the combination of Stable Diffusion and DINOv2 (Oquab et al., 2023) features significantly improves performance in semantic correspondence task. Following (Zhang et al., 2024), we add DINOv2 features to our TED and report the tracking results in Table 9. Table 9 shows that incorporating additional DINOv2 features in our TED does not further improve tracking performance.

Table 9: **TED with additional DINOv2 features**. We introduce additional DINOv2 features as a complementary to our TED method following Zhang et al. (2024). We find that additional DINOv2 features do not further improve the performance of our TED in the tracking task.

Model	Features	Kubric-Similar			Youtube-Similar			DAVIS		
		$\mathcal{J}\&\mathcal{F}_m(\uparrow)$	$\mathcal{J}_m(\uparrow)$	$\mathcal{F}_m(\uparrow)$	$\mathcal{J}\&\mathcal{F}_m(\uparrow)$	$\mathcal{J}_m(\uparrow)$	$\mathcal{F}_m(\uparrow)$	$\mathcal{J}\&\mathcal{F}_m(\uparrow)$	$\mathcal{J}_m(\uparrow)$	$\mathcal{F}_m(\uparrow)$
TED (With DINOv2 features)	ADM, I2VGen-XL, DINOv2	90.0	86.6	93.5	65.9	65.0	66.7	77.3	74.2	80.5
TED (Ours)	ADM, I2VGen-XL	**90.4**	**86.9**	**94.0**	**66.0**	**65.1**	**67.0**	**77.6**	**74.4**	**80.8**