Video Diffusion Models Learn the *Structure* of the *Dynamic* World

Zhipeng Bao[1] Anurag Bagchi[1] Yu-Xiong Wang[2] Pavel Tokmakov[3] Martial Hebert[1]
[1]Carnegie Mellon University [2]University of Illinois Urbana-Champaign [3]Toyota Research Institute
{zbao, anuragba, hebert}@cs.cmu.edu pavel.tokmakov@tri.global yxw@illinois.edu

Abstract

Diffusion models have demonstrated remarkable progress in visual perception tasks due to their ability to capture fine-grained, object-centric features. In this work, we investigate their potential for video understanding by analyzing feature representations learned by both image- and video-based diffusion models, alongside non-generative, self-supervised approaches. To this end, we propose a unified probing framework to evaluate eight models across four core video understanding tasks: action recognition, point tracking, object discovery, and label propagation. Our findings reveal that video diffusion models consistently rank among the top performers, particularly excelling at modeling temporal dynamics and scene structure. This observation not only sets them apart from image-based diffusion models but also opens a new direction for advancing video understanding, offering a fresh alternative to traditional discriminative pre-training objectives. Furthermore, we provide practical insights for leveraging video diffusion models in video understanding, including optimal layer selection, effective fine-tuning strategies, and proxy task-based checkpoint selection. Overall, our results highlight the promise of video diffusion models in capturing both spatial and temporal information, positioning them as strong contenders in video understanding.

1. Introduction

Beyond generating high-fidelity images, diffusion models have made significant strides in visual perception. Their success is driven by large-scale vision-language pretraining, which enables learning of rich, object-centric representations. This has positioned them as strong contenders for tasks such as image segmentation [82, 88] and classification [40]. Given their effectiveness in image-based tasks, a natural question arises: *Can the strengths of diffusion models extend to the more complex domain of videos?*

Video understanding requires not only discerning objects in individual frames but also capturing the complex temporal dynamics that unfold over time. Video diffusion models are architecturally designed to meet these challenges. By integrating spatio-temporal information – using techniques such as temporal attention and 3D convolutions – these models naturally encode both motion trajectories and evolving scene structures. For instance, as shown in Figure 1, when several representations are visualized via K-Means clustering and three-channel PCA, the features extracted by video diffusion models exhibit much stronger temporal consistency, compared to those of image-based models. Moreover, by preserving object-centric representation, these models enhance the implicit understanding of object relationships and environmental context. This dual ability to simultaneously model both *motion* and *structure* underlines the unique benefits of video diffusion models.

To quantitatively study the effectiveness of these models in video understanding, we introduce a unified probing framework across a range of video understanding tasks. This framework systematically dissects the strengths and weaknesses of different visual representations, offering actionable guidance on their optimal use. To ensure a comprehensive analysis, our evaluation spans eight models, including both image- and video-based architectures, as well as non-diffusion approaches. In the diffusion category, we further evaluate both UNet-based [5, 60, 75] and DiT-based models [14, 53].

Our study focuses on four key tasks – two supervised and two unsupervised ones – that highlight different aspects of video understanding: (1) *action recognition*, a supervised classification task for assessing global video-level representations; (2) *point tracking*, a supervised task measuring dense feature quality and motion capturing capability; (3) *object discovery*, a self-supervised task to test the object permanence of the features; and (4) *label propagation*, a training-free task evaluating the temporal consistency of features. Together, these tasks provide a comprehensive analysis of each model's strengths and limitations.

Key insights from our study include:

- Video diffusion models excel at capturing motion dynamics while maintaining a high-level understanding of the structure of the visual world, which supports their consistently strong performance.

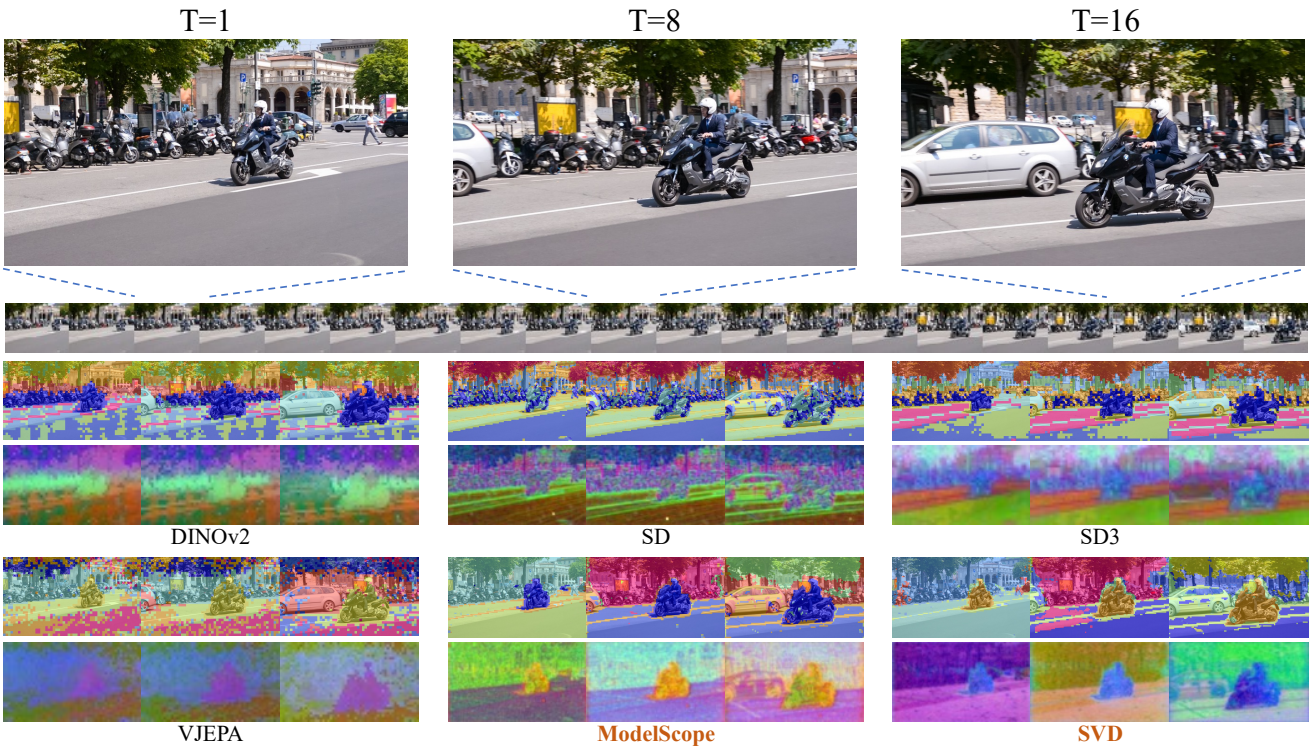

Figure 1. Video feature visualizations on DAVIS17 [57] dataset, where video diffusion models are marked in **orange**. Row 1: K-Means clusters (K=10); Row 2: three-channel PCA visualizations. Compared to image diffusion, or discriminatively trained models, video diffusion models excel at capturing motion dynamics while retaining a higher-level structured representation of the video input. These unique characteristics position them as strong candidates for video understanding.

- These models encode different information at various layers: early layers focus on abstract, high-level features, while later layers capture finer details.
- Progress in video generation does not always correlate with performance on downstream tasks. However, proxy tasks can help systematically identify the best-performing checkpoint.

Overall, video diffusion models show promise for video understanding, excelling at capturing the dynamic structure of the visual world via Web-scale generative pre-training.

2. Related Work

Diffusion models. Inspired by principles of heat and anisotropic diffusion, diffusion models have emerged as a powerful class of generative models for image and video synthesis [55, 79]. Recent advancements have positioned diffusion models as state-of-the-art across unconditional [7, 12, 27, 67, 68] and conditional image synthesis tasks [19, 26, 50, 59, 60, 63, 76, 83, 86]. Notably, Denoising Diffusion Probabilistic Models (DDPMs) [27] introduced the use of neural networks for modeling the denoising process, optimizing with a weighted variational bound. The Denoising Diffusion Implicit Model (DDIM) [27] enhanced this by incorporating a non-Markov sampling strategy to accelerate inference. Stable Diffusion [60] extended the diffusion-denoising process into the latent space of a pre-trained autoencoder [37], enabling more efficient large-scale model training. More recently, Transformer-based models have been introduced to further scale up training, achieving superior performance [14, 53].

The extension of diffusion models from image to video generation [23, 29, 44] gains remarkable achievements, encompassing both text-to-video (T2V)[6, 33, 35, 58, 77] and image-to-video (I2V) generation[21, 49, 74, 87].These efforts largely build upon pre-trained image-level diffusion models, such as Stable Diffusion [60], by training the additional video backbone with extra video data [5, 9, 10, 16, 22, 28, 75]. Some approaches avoid retraining entirely by utilizing training-free algorithms for video generation from image models [66, 81, 84]. Among them, ModelscopeT2V [75] and Stable Video Diffusion (SVD) [5] have open-sourced their large-scale pre-trained model which serves as our backbones for this study.

Diffusion models for visual perception. Diffusion models have also demonstrated strong semantic correspondence in their feature spaces [25, 70, 85]. This has spurred a line of research that utilizes diffusion models for visual percep-

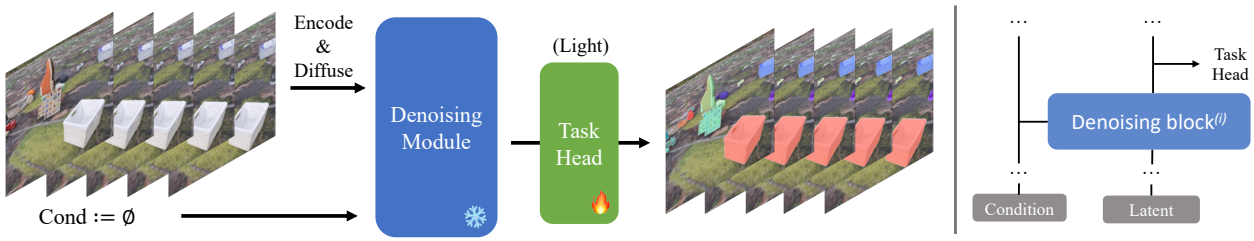

Figure 2. The architecture of our probing framework for video understanding using diffusion models. Video feature representations are extracted from the denoising module, followed by a lightweight task head to produce task-specific annotations. Diffusion representations, intermediate features from the denoising module, are demonstrated on the right.

tual tasks, through either training diffusion-based models for specific tasks such as segmentation [52, 82, 88], depth estimation [20, 64, 65] or open-world novel view synthesis [41]. Other work leverages pre-trained *frozen* diffusion models for perceptual learning [24, 36, 43, 48, 70, 85], or explores their use in data augmentation for discriminative tasks [8, 15, 46, 72].

Among them, DIFT [70] proposes a general pipeline to extract features from real images with diffusion models, which we adopt in our evaluation pipeline. Chen et al. [11] and Nag et al. [47] leverage diffusion models for video-related tasks, but they *do not* leverage a video diffusion model with spatial-temporal reasoning modules. GenRec [80] proposes a joint optimization for video generation and recognition to better facilitate the learning of each other. VD-IT [89] and REM [1] leverage video diffusion models specifically for referring object segmentation. Lexicon3D [45] conducted a comprehensive study of visual foundation models, including diffusion-based ones, on 3D scene understanding. Unlike previous work, this study addresses the general video understanding with diffusion models across multiple tasks, each with a distinct focus.

3. Probing Video Understanding with Diffusion Models

3.1. Preliminaries

Latent diffusion models. Diffusion models [27] are latent variable models that learn the data distribution with the inverse of a Markov noise process. Latent diffusion models (LDM) [60] further switch the diffusion-denoising mechanism from RGB space to latent space, which improves the scalability and enables large-scale training. Concretely, an encoder $\mathcal{E}$ is trained to map a given image $x \in \mathcal{X}$ into a spatial latent code $z = \mathcal{E}(x)$. A decoder $\mathcal{D}$ is then tasked with reconstructing the input image such that $\mathcal{D}(\mathcal{E}(x)) \approx x$.

Considering the clean latent $z_0 \sim q(z_0)$, where $q(z_0)$ is the posterior distribution of z_0, LDM gradually adds Gaussian noise to z_0 in the *diffusion process*:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

The denoising process takes inverse operations from the diffusion. The denoised latent at timestep $t-1$ is estimated via:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \mathbf{\Sigma}_\theta(z_t, t)), \tag{2}$$

where the parameters $\mu_\theta(z_t, t), \mathbf{\Sigma}_\theta(z_t, t)$ of the Gaussian distribution are learned by the denoising network Σ_θ. As shown in [27], $\mathbf{\Sigma}_\theta(z_t, t)$ has only a marginal effect on the results, therefore estimating $\mu_\theta(z_t, t)$ becomes the main objective. A reparameterization is introduced to estimate it:

$$\mu_\theta(z_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t)\right), \tag{3}$$

where $\epsilon_\theta(z_t, t)$ is typically a denoising UNet module [61] or diffusion transformer [53] module. $\epsilon_\theta(z_t, t)$ is usually conditioned on additional inputs, such as texts or image embeddings, to steer the denoising trajectory. In Figure 2 (left), we demonstrate how the extra modality is fused to the latent space: for UNet-based models, cross-attention modules are utilized to fuse the features while for DiT-based models, the additional embedding is fused via AdaIn [31] modules together with the broadcasted self-attention. The final objective of latent diffusion models is:

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]. \tag{4}$$

Video diffusion models generally share a similar architecture to the 2D diffusion models. Given a video $\mathbf{v} = [x^1, x^2, \cdots, x^N]$, a spatial encoder $\mathcal{E}^v$ is applied to each frame to map them to the latent code $z^i = \mathcal{E}^v(x^i)$, where i is the frame index. We use the notation $\mathbf{z} = [z^1, z^2, \cdots, z^N]$ for convenience. For the decoder, usually, a spatio-temporal decoder is applied to enforce the temporal consistency $\mathcal{D}^v(\mathbf{z}) \approx \mathbf{v}$.

One crucial distinction for video diffusion models is that they explicitly model spatio-temporal information with the denoising network, denoted as ϵ_θ^v. This network is extended

to 3D by either introducing additional temporal attention modules [5, 73], or replacing the spatial attention modules with spatio-temporal ones [75, 87].

3.2. Video Understanding Probing Framework

Figure 2 illustrates our unified probing framework. We extract video representations from the denoising module and subsequently apply a lightweight task-specific head for various tasks.

3.2.1. Diffusion Features

We extract video features with diffusion models following DIFT [70]. The process begins by adding noise at timestep T to the real video latent (Equation 1), moving it into the $\mathbf{z}_T$ distribution. This noisy video latent, along with T, is then passed to ϵ_θ^v. Instead of using the final output of ϵ_θ^v, which predicts the noise, we extract features from intermediate layer activations that effectively capture the video's underlying representations:

$$\mathbf{z}_{\text{feature}} = \epsilon_\theta^{v(n)}(\mathbf{z}_T, T), \tag{5}$$

where (n) indicates the block index. Following DIFT, we extract the intermediate representations from upsampling blocks, forming the diffusion features. For features from image diffusion models, we follow a nearly identical process, except that we process the videos frame by frame. Additionally, during feature extraction, we introduce a fixed "null-embedding" as the condition for ϵ_θ^v. For language-based models, this embedding is obtained by passing an empty prompt to the text encoder. For image-based models, we use an all-zero conditional image.

3.2.2. Adaptation for Downstream Tasks

After extracting features from diffusion models, we employ a lightweight task head –fewer than 1% of the backbone's parameters – to adapt these features for the target tasks, as demonstrated by the object discovery task in Figure 2. Below, we detail the task-specific adaptations used in our evaluation.

Action recognition involves predicting a one-hot action label for a given video. Following prior work [4, 71], we first average the extracted feature map and then apply a two-layer MLP, where the hidden dimension matches the input feature size, to produce the final classification output.

Point tracking aims to predict the trajectory of query points over time. Given a spatiotemporal coordinate (t, y, x), where t denotes the frame index and (x, y) represents the spatial location, the model processes a video clip of T frames ($T \geq t$) and outputs a sequence of T points corresponding to the query's trajectory. We employ TAPNet [13] as the backbone, using frozen representations from foundation models to predict the trajectory.

Model	Type	Architecture	Dataset	Feature Dim	Downsample
DINOv2 [51]	Image	ViT-L	LVD-142M	1024	14
VideoMAE [71]	Video	ViT-L	Kinetics400-240k	1024	16
VJEPA [4]	Video	ViT-L	VideoMix2M	1024	16
IV2-1B [78]	Video	InternViT	InternVideo-402M	1408	16
SD [60]	Image	UNet	LAION-5B	1280/640	8/16
SD3 [14]	Image	DiT	PublicImgs-1B	1536	16
ModelScope [75]	Video	UNet	WebVid-10M	1280/640	8/16
SVD [5]	Video	UNet	LVD-152M	1280/640	8/16

Table 1. Details of the pretrained visual foundation models we used for our video understanding evaluation.

The task of **Object discovery** seeks to identify and track dynamic objects in videos in a self-supervised manner. We adopt MoTok [2] for this task, utilizing cross-attention layers with learnable queries, referred to as slots [42], to segment and group foreground regions into distinct objects. Feature-level reconstruction serves as the supervisory signal, with the extracted representations functioning as the reconstruction space.

Label propagation is a training-free task that transfers instance masks or keypoints from an initial frame to subsequent frames, relying on temporal appearance consistency. Instead of predicting new labels, the task propagates existing ones frame-by-frame. Following previous methods [32, 70], we implement this using a k-nearest neighbors (k-NN) search over a feature queue containing the initial frame and the most recent m frames, eliminating the need for a dedicated task head.

4. Experimental Evaluations

4.1. Evaluation Settings

Baseline Models. We perform our video understanding analysis with eight visual foundation models. **DINOv2** [51] is a contrastive learning-based image-level foundation model. **VJEPA** [4] and **VideoMAE** [71] learn comprehensive video representations by reconstructing from masked video patches. **InternVideo2 (IV2-1B)** [78] learns a large-scale video foundation model using video-language contrastive learning. We use their 1B variant, which best match the size of the diffusion models, for our evaluation. **Stable Diffusion (SD)**[60] and **Stable Diffusion 3 (SD3)**[14] are text-to-image diffusion model with UNet [61] and DiT [53] as denoising backbones. **ModelScopeT2V** [75] and **Stable Video Diffusion (SVD)** are video diffusion models that take SD as the initialization and further fine-tune on large-scale video data. Detailed configurations of these feature extractors are provided in Table 1.

Datasets and metrics. We evaluate *action recognition* recognition with top 1 and top 5 accuracy on Something-Something v2 (SSv2) [17] and UCF101 [69]. We also include HMDB51 [38] for this task inside our ablation study. We study the point tracking task on TAP-Vid [13] benchmark, and take Average Jaccard (AJ) and average

Backbone	SSv2		UCF101		TAP-Vid-DAVIS		TAP-Vid-RGB		MOVi-E		DAVIS17			JHMDB	
	Top1 Acc	Top 5 Acc	Top1 Acc	Top 5 Acc	AJ	$< \delta^x_{avg}$	AJ	$< \delta^x_{avg}$	FG.ARI	mBO	$\mathcal{J}_m$	$\mathcal{F}_m$	$\mathcal{J}\&\mathcal{F}_m$	PCK@0.1	PCK@0.2
DINOv2	36.5	68.8	89.8	97.8	24.7	36.6	43.7	58.5	71.9	26.3	64.8	69.1	67.0	50.4	78.7
VideoMAE	38.1	71.2	87.9	97.9	12.3	23.0	26.1	38.9	32.7	14.1	30.5	37.5	34.0	32.5	59.3
VJEPA	42.2	78.4	92.1	98.5	14.6	24.5	28.0	40.6	49.9	18.0	52.3	58.0	55.1	37.5	70.3
IV2-1B	46.3	81.3	94.0	98.9	17.8	26.1	30.1	41.6	47.2	18.3	33.0	38.6	35.8	39.2	70.1
SD	24.7	52.1	63.5	86.1	30.8	47.4	46.5	63.8	63.4	26.9	67.8	74.6	71.2	60.5	80.8
SD3	21.9	48.7	60.9	85.8	23.8	35.6	44.5	58.2	65.1	28.6	48.5	54.8	51.6	38.2	65.9
ModelScope	41.4	75.3	80.6	94.9	39.5	57.4	57.0	73.1	63.7	27.5	65.3	72.4	68.4	60.9	82.8
SVD	43.0	79.2	92.3	98.6	36.9	51.9	53.1	68.8	65.4	29.4	59.8	67.7	63.8	60.5	81.8

Table 2. Quantitative evaluations on the four evaluated tasks. The top two results are marked in green and yellow respectively. Video diffusion models provide semantic- and geometric-aware representations that contain both high-level abstractions and detailed information, positioning them as unique and competitive candidates for video understanding.

positional accuracy under multi-scale thresholds ($< \delta^x_{avg}$) as metrics. We evaluate the object discovery on MOVi-E [18], and use foreground adjusted random index (FG. ARI) and video mean best overlap (mBO) as metrics. We conduct the label propagation for video object segmentation on DAVIS17 [57] and keypoints estimation on JHMDB [34] following the same setup as DIFT [70]. We report region-based similarity $\mathcal{J}$ and contour-based accuracy $\mathcal{F}$ [54] for DAVIS17, and percentage of correct keypoints (PCK) for JHMDB.

Key implementation details. We use the noise level 50 by default, with a corresponding timestep T=50 (for SD, ModelScope, and SVD) or T=16 (for SD3). For the layer index, we design the use of block index 1 (for SD, ModelScope, and SVD) and layer index 12 (for SD3) for action recognition. For the other tasks, we use block index 2 and layer index 24 respectively. We select these configurations based on the observations we obtained from Section 4.3. We use batch size 12 with 4 NVIDIA-A100 GPUs running in parallel for all the backbones except ModelScope. We use batch size 6 with 8 GPUs in parallel for ModelScipe to fit its CUDA requirement.

More details about datasets, model implementation, and training configurations are included in Section C of the supplementary material.

4.2. Main Results

We present the quantitative results of our evaluation in Table 2 and representative visual comparisons in Figure 3. Additional results, including comparisons with state-of-the-art methods and evaluations with stronger task heads, are provided in Section B.4 of the supplementary material.

For the following discussions, we treat ModelScope and SVD as variants of the "video diffusion model" category, despite differences in conditioning (Text-to-Video *vs*. Image-to-Video). Given the lack of a standardized training, we focus on their shared foundations instead: both models are based on SD with additional video training, which enables us to study their common strengths and limitations.

Overall conclusions. Most representations only exhibit strong performance on individual tasks. For instance, contrastive learning approaches like VJPEA or IV2-1B excel in global classification task, but perform very poorly in dense prediction. Similarly, image-based models predictably with motion understanding. In contrast, video diffusion models consistently rank among the top performers across all tasks and datasets, highlighting their robustness and adaptability in video understanding. As shown in Figure 3, video diffusion models effectively capture motion dynamics while also providing a more structured representation of the visual world compared to both image-based diffusion models and discriminative video models.

Action recognition. Surprisingly, SVD achieves the second best performance on both datasets, only lagging behind the state-of-the-art, video-language IV2-1B model and consistently outperforming representations that are specifically tuned for action recognition, like VJEPA and VideoMAE. This result highlights the ability of well-trained video diffusion models to capture global-level video representations effectively. Interestingly, SD3, which uses a DiT-based architecture, performs suboptimally. A possible explanation lies in how DiT models fuse multi-modal features, suggesting an open research direction: improving feature extraction techniques tailored for DiT-based diffusion models.

Point tracking. Both video diffusion models, ModelScope and SVD emerge as the top performers in this task, outperforming other baselines by a large margin. As shown in Figures 1 and 3, the key factor behind their success is their superior capacity to capture motion, which plays a key role in point tracking.

Object discovery. DINOv2, known for its strong object segmentation capabilities, performs competitively on this object-centric task. However, SVD outperforms it in terms of mBO, which emphasizes distinguishing objects from the background and *precisely* tracking them. This suggests that diffusion models are particularly well-suited for tasks requiring fine-grained localization and tracking. Visual comparisons in Figure 3 provide further evidence where SVD precisely tracks objects with complex motion.

Label propagation. On DAVIS17, video diffusion models lag behind image-level StableDiffusion. We hypothesize that this is because video diffusion models learn detailed representations of moving objects (refer to Figure 1) but struggle to differentiate static objects from the background

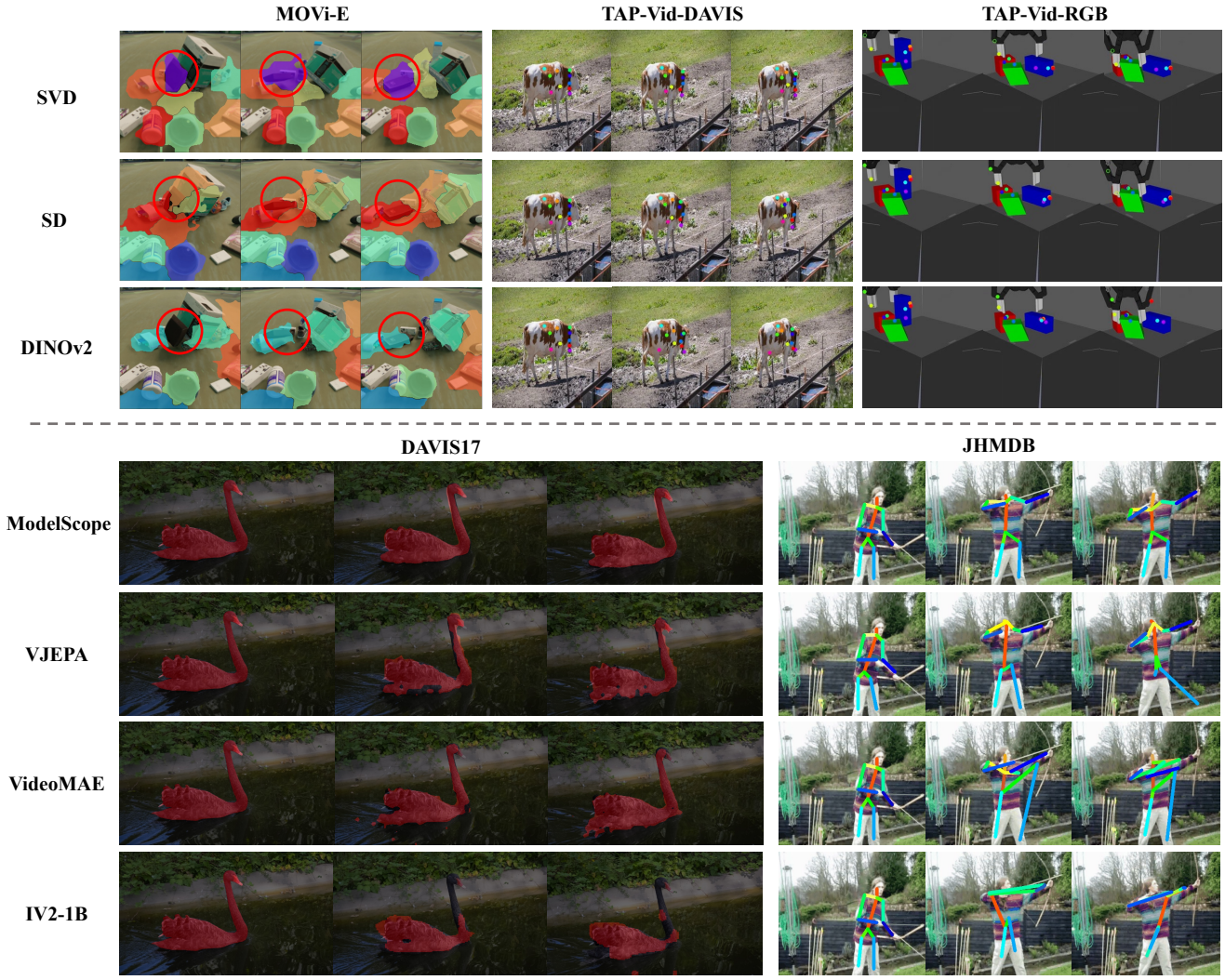

Figure 3. Visual comparisons between the results of video diffusion models and other foundation models. **Top:** Video diffusion models capture motion dynamics more effectively than image-based models; **Bottom:** Video diffusion models demonstrate a stronger understanding of world structure compared to conventional video foundation models. This balance of dynamic and structural comprehension enables them to consistently perform at a high level. Video versions and more visualizations are provided in the supplementary demo video.

- a key challenge in video object segmentation (VOS). In contrast, on the JHMDB dataset, where keypoints estimation focuses solely on a single moving object, video diffusion models demonstrate their strengths.

In summary, video diffusion models provide semantic- and geometric-aware representations that effectively capture both high-level abstractions and detailed information, positioning them as versatile representations for video understanding.

4.3. Guidelines for Video Diffusion Adoption

In our main evaluation, we use frozen video diffusion representations with fixed noise levels and predefined layer indices. In this section, we investigate how to better utilize these representations by providing guidance on layer selection, fine-tuning strategies, and model selection. For action recognition experiments we utilize the small-scale HMDB51 [38] dataset to reduce the computational costs.

Noise levels and block indices. We examine the effects of noise levels and block indices in SVD across three tasks: action recognition on HMDB51, object discovery on MOVi-E, and label propagation on DAVIS17, as summarized in Table 3. We omit point tracking since it explicitly requires grid sampling from the feature map, where later blocks (with higher resolution) naturally provide an advantage. Our results suggest that noise level plays a relatively minor and task-specific role compared to block indices. Generally, low noise levels (*e.g.*, corresponding to $T = 50$) yields strong

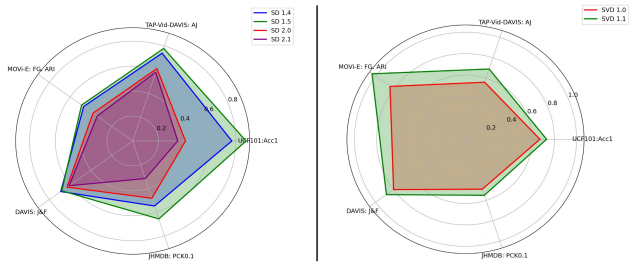

Figure 4. Comparison between newer and older versions of the same model series. The newer version of the same model is not always better for video understanding, while the task performances within the same model series are highly correlated – a single version consistently emerges as the strongest across all tasks. This observation suggests that a proxy task can serve as an effective alternative for identifying the optimal model version.

Noise Level	Block Index	HMDB51		MOVi-E		DAVIS17		
		Top1 Acc	Top5 Acc	FG.ARI	mBO	$\mathcal{J}_m$	$\mathcal{F}_m$	$\mathcal{J}\&\mathcal{F}_m$
0	1	60.3	88.0	58.1	23.4	52.1	44.9	48.5
50	1	63.8	**89.7**	58.3	23.7	51.1	42.6	46.9
100	1	**63.9**	89.4	57.9	22.8	50.3	41.6	46.0
200	1	62.6	88.7	57.6	23.5	50.2	41.3	45.8
0	2	31.1	64.0	**65.8**	**29.4**	**60.8**	**68.0**	**64.4**
50	2	33.7	66.9	65.4	**29.4**	59.8	67.7	63.8
100	2	35.4	68.0	64.3	28.5	59.6	67.2	63.4
200	2	32.8	66.8	63.9	27.9	59.1	64.5	62.8

Table 3. Ablation on noise levels and block indices of SVD. Compared to noise level, the block index has a significant impact on downstream task performance. Features from earlier blocks capture more abstract, high-level information, while features from later blocks are more object-oriented.

results. In contrast, block indices significantly influence downstream task performance: features from earlier blocks encode abstract, high-level information, making them ideal for classification tasks, while features from later blocks capture finer details, benefiting dense prediction tasks. Additional feature visualizations of block index effects are provided in Section B.1 of the supplementary. These observations align with findings from image diffusion models [70].

Finetuning video diffusion models. For certain video understanding tasks, fine-tuning the backbone is essential and typically results in improved performance. To explore the impact and strategies of fine-tuning video diffusion models for perception tasks, we fine-tune the SVD denoising UNet on HMDB51 and MOVi-E. The results are summarized in Table 4 (first two rows). For object discovery, we slightly modify the baseline architecture of Bao et al. [2], with details provided in Section C.4 in the supplementary material. As a result, the reported FG.ARI score for the frozen model differs from that in Table 2.

Notably, by comparing the change of parameters of all the modules, we find that the last in-use block (*e.g.* block index 1 for action recognition and block index 2 for object discovery) exhibits the highest sensitivity to parameter changes (details in the supplementary material, Section B.2), highlighting their critical role in enhancing task performance. Inspired by previous efficient diffusion finetuning approaches [3, 39, 62], we construct two fine-tuning variants: one incorporates LoRA [30] adaptation layers in all attention blocks, while the other fine-tune only the last block in use, the most parameter-sensitive block. The results for these two variants are reported in Table 4 (last two rows). These findings demonstrate that efficient finetuning strategies can significantly enhance performance while keeping training costs reasonable, offering practical guidance for optimizing video diffusion models.

Strategy	HMDB51			MOVi-E		
	Top1 Acc	Mem.	Time	FG.ARI	Mem.	Time
Frozen	63.8	1.0 ×	1.0 ×	66.1	1.0 ×	1.0 ×
Full	68.3	2.6 ×	2.3 ×	69.2	2.7 ×	2.5 ×
LoRA	66.9	1.1 ×	1.7 ×	67.0	1.2 ×	1.7×
Last-in-use	67.1	1.3 ×	1.8 ×	68.1	1.4 ×	1.9×

Table 4. Performance and training cost for finetuning SVD UNet. "Last-in-use" denotes only fine-tuning the last in-use upsampling block. While finetuning the diffusion backbone yields performance improvements, it comes with significantly higher computational costs. Using an efficient finetuning strategy by only tweaking the most sensitive layers leads to an effective.

Model selection. Modern diffusion models are continuously evolving, with newer versions improving in generation quality. However, whether newer versions offer better visual perception capabilities remains unclear. This raises an important question: *does a newer version of the same model inherently perform better in visual perception tasks?* To investigate this, we systematically evaluate four versions of Stable Diffusion (SD v1.4, 1.5, 2.0, and 2.1) and two versions of Stable Video Diffusion (SVD v1.0 and v1.1) across our four core video understanding tasks. The results, summarized in Figure 4, reveal an intriguing trend: newer is not always better for video understanding. While some later versions (*e.g.*, SD 1.5) outperform their predecessors across all tasks, others (*e.g.*, SD 2.1) lead to decreased performance.

Interestingly, within the same model series, task performances are highly correlated – a single version consistently emerges as the strongest across all tasks (*e.g.*, SD 1.5 and SVD 1.1). This suggests that model version or generation quality alone is not a reliable criterion for selecting the best representation for video understanding. Instead, a proxy task can serve as an effective alternative for identifying the optimal model. For example, the learning-free label propagation can serve as a proxy for point tracking, providing a simple yet effective heuristic for selecting the best-performing model.

Model	SSv2 Top1 Acc	UCF101 Top1 Acc	TAP-Vid-DAVIS AJ	TAP-Vid-RGB AJ
SD	24.7	63.5	30.8	46.5
SDXL	26.9	70.1	28.5	42.1
ModelScope	41.4	80.6	**39.5**	**57.0**
SVD	**43.0**	**92.3**	36.9	53.1

Table 5. Comparison with a larger image diffusion model – SDXL [56]. SDXL improves over SD in action recognition but underperforms in point tracking. Crucially, both models still lag significantly behind video diffusion models, confirming that the performance gains of video diffusion models stem from additional video training rather than simply increased model capacity.

Modle	MOVi-E		JHMDB	
	FG.ARI	mBO	PCK@0.1	PCK@0.2
DINOv2	**68.3**	25.9	51.3	79.0
VideoMAE	29.4	13.7	30.2	57.1
VJEPA	49.0	18.2	37.7	69.8
SD	63.4	26.9	60.5	80.8
SD3	65.9	29.1	42.4	73.8
ModelScope	63.7	27.5	**60.9**	**82.8**
SVD	65.4	**29.4**	60.5	81.8

Table 6. Evaluation under the same feature resolution. With a higher input resolution, only the performance of SD3 gets consistently improved. However, the overall conclusions regarding the effectiveness and robustness of video diffusion models holds.

4.4. Discussion

Effect of the model size. Both ModelScope and SVD introduce additional parameters compared to SD. To isolate the impact of model size from improvements due to additional video training, we introduce a larger image diffusion model, StableDiffusionSL (SDXL) [56], which contains more parameters than both SVD and ModelScope. We evaluate SDXL on action recognition and point tracking, where SD significantly underperforms compared to video diffusion models, to assess whether increased model size alone can bridge the gap. The results, shown in Table 5, indicate that SDXL improves over SD in action recognition but underperforms in point tracking. Crucially, both models still lag significantly behind video diffusion models, confirming that the performance gains of video diffusion models stem from additional video training rather than simply increased model size.

Evaluation under the same feature resolution. In our main evaluation, we maintain a consistent input video resolution across all models. However, due to varying downsampling ratios, the resulting feature map resolutions differ, potentially affecting performance. To account for this, we conduct an additional evaluation where we upsample input data for models with a downsampling ratio larger than 8, ensuring uniform feature resolution across all models. We exclude action recognition as features are average pooled for this task, and point tracking since it is non-trivial to apply such design under the standard evaluation protocol. The results, presented in Table 6, show that higher input resolution only consistently improves SD3's performance. However, the overall conclusions regarding the effectiveness and robustness of video diffusion models remain unchanged.

Figure 5. Limitations of Video Diffusion Representations: difficulty in handling occlusion among instances of the same semantic category, and challenges with distinguishing nearby objects that share similar motion.

Limitations of video diffusion representations. Despite their strong performance, video diffusion representations exhibit certain limitations. We specifically analyze failure cases in label propagation because it does not rely on a trainable task head, making the observations more direct and reflective of the underlying representations. As illustrated in Figure 5, two common failure modes emerge: (1) difficulty handling occlusion among instances of the same semantic category and (2) challenges in distinguishing nearby objects with similar motion patterns. These limitations highlight potential areas for future improvements in video diffusion-based representations.

5. Conclusion and Future Work

This paper showcases that video diffusion models offer a powerful approach to video understanding, excelling in capturing motion dynamics and high-level structural representations. By systematically analyzing their performance across multiple tasks, we highlight their robustness, adaptability, and the distinct advantages they bring to video perception. These models stand out for their unique balance of dynamic and structural comprehension, positioning them as promising tools for advancing video understanding. Moreover, our findings provide actionable insights into how their representations can be optimized through careful layer selection and fine-tuning strategies, paving the way for more efficient and effective utilization of video diffusion models in various applications.

Promising directions for **future work** include: (1) designing a more advanced feature extraction pipeline with newly introduced DiT-based models; (2) further exploring techniques to better leverage video diffusion representations, including custom head design and deep fine-tuning [1, 80].

References

[1] Anurag Bagchi, Zhipeng Bao, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Refereverything: Towards segmenting everything we can speak of in videos. *arXiv preprint arXiv:2410.23287*, 2024. 3, 8

[2] Zhipeng Bao, Pavel Tokmakov, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Object discovery from motion-guided tokens. In *CVPR*, 2023. 4, 7

[3] Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. Separate-and-enhance: Compositional finetuning for text-to-image diffusion models. In *SIGGRAPH*, 2024. 7

[4] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024. 4

[5] 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. 1, 2, 4

[6] 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. 2

[7] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *ECCV*, 2022. 2

[8] Max F. Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. A data augmentation perspective on diffusion models and retrieval. *arXiv preprint arXiv:2304.10253*, 2023. 3

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[10] 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. 2

[11] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *ICCV*, 2023. 3

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 2

[13] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. TAP-vid: A benchmark for tracking any point in a video. *NeurIPS*, 2022. 4

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 4

[15] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *ICCV*, 2023. 3

[16] 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. 2

[17] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 4

[18] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. In *CVPR*, 2022. 5

[19] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 2

[20] Vitor Guizilini, Pavel Tokmakov, Achal Dave, and Rares Ambrus. Grin: Zero-shot metric depth with pixel-level diffusion. *arXiv preprint arXiv:2409.09896*, 2024. 3

[21] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Chongyang Ma, Weiming Hu, Zhengjun Zha, Haibin Huang, Pengfei Wan, et al. I2v-adapter: A general image-to-video adapter for video diffusion models. *arXiv preprint arXiv:2312.16693*, 2023. 2

[22] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[23] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2

[24] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *NeurIPS*, 2023. 3

[25] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2, 3

[28] 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. 2

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 7

[31] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3

[32] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *NeurIPS*, 2020. 4

[33] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. *arXiv preprint arXiv:2312.07509*, 2023. 2

[34] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013. 5

[35] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2

[36] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me. *arXiv preprint arXiv:2309.03179*, 2023. 3

[37] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[38] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 4, 6

[39] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 7

[40] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, 2023. 1

[41] Ruoshi Liu, Rundi We, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 3

[42] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, 2020. 4

[43] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *NeurIPS*, 2023. 3

[44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 2

[45] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *arXiv preprint arXiv:2409.03757*, 2024. 3

[46] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[47] Sauradip Nag, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Difftad: Temporal action detection with proposal denoising diffusion. *arXiv preprint arXiv:2303.14863*, 2023. 3

[48] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. *arXiv preprint arXiv:2401.11739*, 2024. 3

[49] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, 2023. 2

[50] 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. 2

[51] 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. *TMLR*, 2023. 4

[52] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *CVPR*, 2024. 3

[53] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1, 2, 3, 4

[54] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 5

[55] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *PAMI*, 1990. 2

[56] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 8

[57] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2, 5

[58] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483*, 2023. 2

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

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4

[61] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3, 4

[62] 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 *CVPR*, 2023. 7

[63] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[64] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In *NeurIPS*, 2023. 3

[65] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 3

[66] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[67] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[68] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[69] 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. 4

[70] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *NeurIPS*, 2023. 2, 3, 4, 5, 7

[71] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 4

[72] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *ICLR*, 2024. 3

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4

[74] Cong Wang, Jiaxi Gu, Panwen Hu, Songcen Xu, Hang Xu, and Xiaodan Liang. Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance. *arXiv preprint arXiv:2312.03018*, 2023. 2

[75] 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. 1, 2, 4

[76] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 2

[77] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. *arXiv preprint arXiv:2312.15770*, 2023. 2

[78] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 2024. 4

[79] Joachim Weickert et al. *Anisotropic diffusion in image processing*. Teubner Stuttgart, 1998. 2

[80] Zejia Weng, Xitong Yang, Zhen Xing, Zuxuan Wu, and Yu-Gang Jiang. Genrec: Unifying video generation and recognition with diffusion models. *arXiv preprint arXiv:2408.15241*, 2024. 3, 8

[81] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 2

[82] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. ODISE: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 1, 3

[83] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[84] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. *arXiv preprint arXiv:2312.12490*, 2023. 2

[85] 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. *NeurIPS*, 2023. 2, 3

[86] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2

[87] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2, 4

[88] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023. 1, 3

[89] Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan, Chunming Qiao, and Gang Hua. Exploring pre-trained text-to-video diffusion models for referring video object segmentation. *arXiv preprint arXiv:2403.12042*, 2024. 3