

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

Anonymous CVPR submission

Paper ID 9000

Abstract

Diffusion models have demonstrated significant progress in visual perception tasks due to their ability to capture fine-grained, object-centric features. In this work, we explore the potential of diffusion models for video understanding by analyzing the feature representations learned by both image- and video-based diffusion models, alongside non-generative, self-supervised approaches. We propose a unified probing framework to evaluate seven models across four core video understanding tasks: action recognition, object discovery, scene understanding, 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. Interestingly, we demonstrate that higher-generation performance does not always correlate with improved performance in downstream tasks, highlighting the importance of careful representation selection. Overall, our results suggest that video diffusion models hold substantial promise for video understanding by effectively capturing both spatial and temporal information, positioning them as strong competitors in this evolving domain.

1. Introduction

Beyond generating high-fidelity images, diffusion models have achieved significant breakthroughs in visual perception. Their success is largely attributed to the large-scale vision-language pretraining, which allows them to capture detailed, object-centric features, and positions them as strong candidates for tasks such as image segmentation [81, 87] and classification [41]. Naturally, this raises a question: *Can diffusion models' success in images extend to the more complex domain of video understanding?*

Video understanding presents unique challenges absent in the image domain, particularly in capturing *temporal dy-*

namics and motion patterns. Unlike image diffusion models, video diffusion models [5, 75] are inherently designed to capture such spatial-temporal dynamics, making them far better suited for these tasks. As illustrated in Figure 1, where we visualize video representations using K-Means clustering and three-channel PCA for several widely used visual foundation models, video diffusion models excel at capturing motion dynamics – a critical capability that sets them apart from their image-based counterparts. Additionally, they retain a high-level structured representation of the visual world, further enhancing their implicit understanding of object relationships and environmental context. This dual capability of modeling both *motion* and *structure* makes them strong candidates for video understanding tasks.

To further investigate the effectiveness of video diffusion models in video understanding, we introduce a unified probing framework to systematically analyze feature representations from diffusion models across a range of video understanding tasks. This framework enables a detailed examination of the relative strengths and limitations of video diffusion models, providing practical insights for their optimal use. To ensure a comprehensive analysis, our evaluation spans seven models, including both image- and video-based architectures, as well as non-diffusion [4, 52, 71] and diffusion-based approaches. In the diffusion category, we further evaluate both UNet-based [5, 60, 75] and diffusion-transformer-based techniques [15, 54, 88].

Our study focuses on four key tasks that highlight different aspects of video understanding: (1) *action recognition*, a supervised classification task for assessing global video-level representations; (2) *object discovery*, an unsupervised segmentation task measuring dense feature quality; (3) *scene understanding*, a supervised task to test the semantic and geometrical awareness; and (4) *label propagation*, a training-free task evaluating the temporal consistency of features. These tasks provide a comprehensive examination of the strengths and weaknesses of each model across various facets of video understanding.

Key insights from our study include:

- Video diffusion models excel at capturing motion dynamics while maintaining a high-level understanding of the

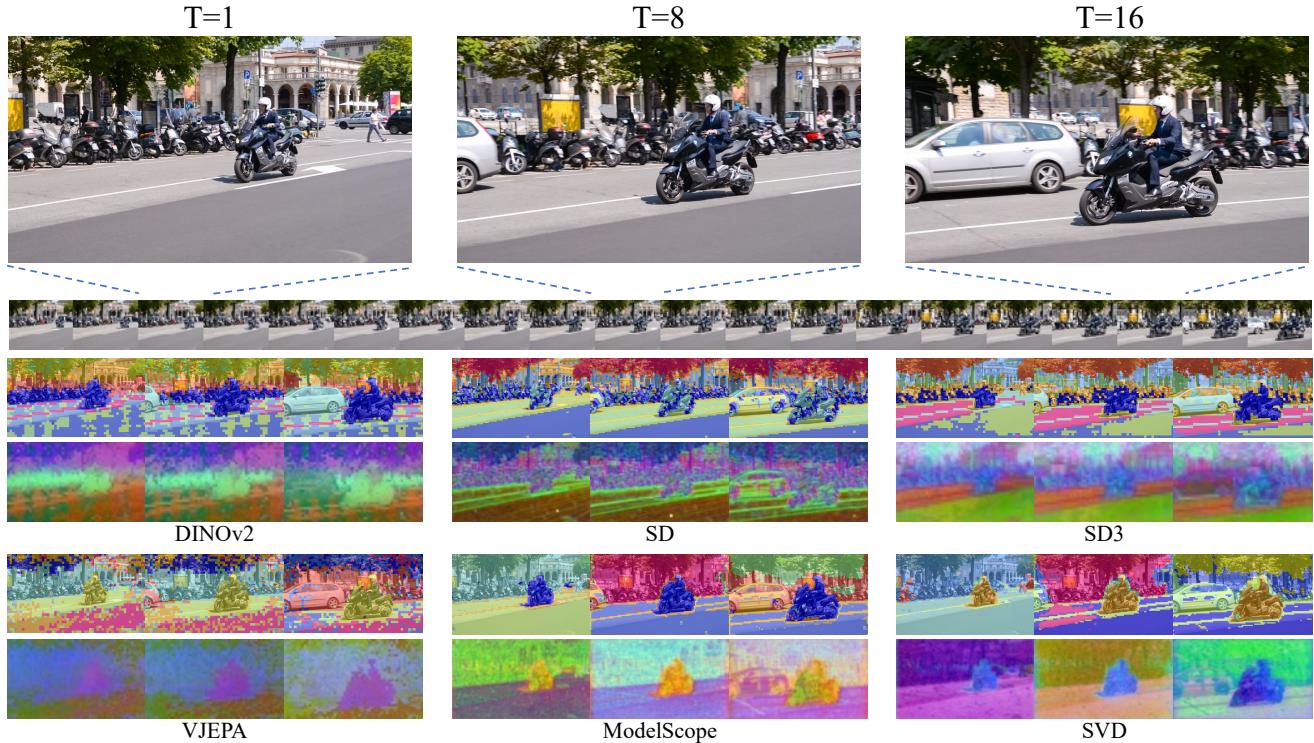


Figure 1. Video feature visualizations on DAVIS17 [57] dataset. Row 1: K-Means clusters ($K=10$); Row 2: three-channel PCA visualizations. Compared to image diffusion, or discriminatively trained models, video diffusion models excel at capturing motion dynamics while retaining a higher-level structured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

079 structure of the visual world, which supports their consistently strong performance.
080

- 081 • These models encode different information at various layers: early layers focus on abstract, high-level features,
082 while later layers capture finer details. Fine-tuning only
083 the most relevant layers enhances adaptation efficiency
084 with minimal performance loss.
- 085 • Surprisingly, greater generative capacity does not always
086 improve performance in visual perception tasks—earlier
087 model versions sometimes outperform newer ones in
088 downstream applications.

089 Overall, video diffusion models show significant promise
090 for video understanding, excelling at capturing the dynamic
091 structure of the visual world and emerging as competitive
092 solutions in this field.

094 2. Related Work

095 **Diffusion Models.** Inspired by principles of heat and
096 anisotropic diffusion, diffusion models have emerged as
097 a powerful class of generative models for image and
098 video synthesis [56, 78]. Recent advancements have
099 positioned diffusion models as state-of-the-art across uncon-
100 ditional [7, 14, 27, 67, 68] and conditional image synthesis

101 tasks [19, 26, 51, 59, 60, 63, 76, 82, 85]. Notably, Denoising
102 Diffusion Probabilistic Models (DDPMs) [27] introduced
103 the use of neural networks for modeling the denoising
104 process, optimizing with a weighted variational bound.
105 The Denoising Diffusion Implicit Model (DDIM) [27] enhanced
106 this by incorporating a non-Markov sampling strategy to
107 accelerate inference. Stable Diffusion [60] extended
108 the diffusion-denoising process into the latent space of a
109 pre-trained autoencoder [37], enabling more efficient large-
110 scale model training. More recently, Transformer-based
111 models have been introduced to further scale up training,
112 achieving superior performance [15, 54].

113 The extension of diffusion models from image to video
114 generation [23, 29, 45] gains remarkable achievements, en-
115 compassing both text-to-video (T2V)[6, 33, 35, 58, 77] and
116 image-to-video (I2V) generation[21, 50, 74, 86]. These ef-
117 forts largely build upon pre-trained image-level diffusion
118 models, such as Stable Diffusion [60], by training the addi-
119 tional video backbone with extra video data [5, 10, 11, 17,
120 22, 28, 75]. Some approaches avoid retraining entirely by
121 utilizing training-free algorithms for video generation from
122 image models [66, 80, 83]. Most recently, Sora [8] and
123 its open-sourced counterparts [40, 88] demonstrated lead-
124 ing video generation capabilities with the more advanced

125 architecture of diffusion transformer [54]. Among them,
 126 ModelscopeT2V [75], Stable Video Diffusion (SVD) [5],
 127 and OpenSora [86] have open-sourced their large-scale pre-
 128 trained model which serves as our backbones for this study.

129 **Diffusion Models for Visual Perception.** Diffusion models
 130 have also demonstrated strong semantic correspondence
 131 in their feature spaces [25, 70, 84]. This has spurred a line
 132 of research that utilizes diffusion models for visual percep-
 133 tual tasks, through either training diffusion-based models
 134 for specific tasks such as segmentation [53, 81, 87], depth
 135 estimation [20, 64, 65] or open-world novel view synthe-
 136 sis [42]. Other work leverages pre-trained *frozen* diffusion
 137 models for perceptual learning [24, 36, 44, 49, 70, 84], or
 138 explores their use in data augmentation for discriminative
 139 tasks [9, 16, 47, 72].

140 Among them, DIFT [70] proposes a general pipeline
 141 to extract features from real images with diffusion mod-
 142 els, which we adopt in our evaluation pipeline. Chen
 143 et al. [12] and Nag et al. [48] leverage diffusion models
 144 for video-related tasks, but they *do not* leverage a video
 145 diffusion model with spatial-temporal reasoning modules.
 146 GenRec [79] proposes a joint optimization for video gen-
 147 eration and recognition to better facilitate the learning of
 148 each other. VD-IT [89] and REM [1] leverage video diffu-
 149 sion models specifically for referring object segmentation.
 150 Lexicon3D [46] conducted a comprehensive study of vi-
 151 sual foundation models, including diffusion-based ones, on
 152 3D scene understanding. Unlike previous work, this study
 153 addresses the general video understanding with diffusion
 154 models across multiple tasks, each with a distinct focus.

155 3. Probing Video Understanding with Diffu- 156 sion Models

157 3.1. Preliminaries

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

167 Considering the clean latent $z_0 \sim q(z_0)$, where $q(z_0)$ is
 168 the posterior distribution of z_0 , LDM gradually adds Gaus-
 169 sian noise to z_0 in the *diffusion process*:

$$170 q(z_t | z_{t-1}) = \mathcal{N}(z_t; \sqrt{1 - \beta_t} z_{t-1}, \beta_t \mathbf{I}), \quad (1)$$

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

$$173 p_\theta(z_{t-1} | z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)), \quad (2)$$

175 where the parameters $\mu_\theta(z_t, t), \Sigma_\theta(z_t, t)$ of the Gaussian
 176 distribution are learned by the denoising network Σ_θ . As
 177 shown in [27], $\Sigma_\theta(z_t, t)$ has only a marginal effect on the
 178 results, therefore estimating $\mu_\theta(z_t, t)$ becomes the main ob-
 179 jective. A reparameterization is introduced to estimate it:

$$180 \mu_\theta(z_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(z_t, t) \right), \quad (3)$$

182 where $\epsilon_\theta(z_t, t)$ is typically a denoising UNet module [61]
 183 or diffusion transformer [54] module. $\epsilon_\theta(z_t, t)$ is usually
 184 conditioned on additional inputs, such as texts or image em-
 185 beddings, to steer the denoising trajectory. In Figure 2 (left),
 186 we demonstrate how the extra modality is fused to the latent
 187 space: for UNet-based models, cross-attention modules are
 188 utilized to fuse the features while for DiT-based models, the
 189 additional embedding is fused via AdaIn [31] modules to-
 190 gether with the broadcasted self-attention. The final objec-
 191 tive of latent diffusion models is:

$$192 \mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0, 1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2 \right]. \quad (4)$$

193 **Video Diffusion Models.** Video diffusion models gener-
 194 ally share a similar architecture to the 2D diffusion mod-
 195 els. Given a video $\mathbf{v} = [x^1, x^2, \dots, x^N]$, a spatial encoder
 196 \mathcal{E}^v is applied to each frame to map them to the latent code
 197 $z^i = \mathcal{E}^v(x^i)$, where i is the frame index. We use the nota-
 198 tion $\mathbf{z} = [z^1, z^2, \dots, z^N]$ for convenience. For the decoder,
 199 usually, a spatio-temporal decoder is applied to enforce the
 200 temporal consistency $\mathcal{D}^v(\mathbf{z}) \approx \mathbf{v}$.

201 One crucial distinction for video diffusion models is that
 202 they explicitly model spatio-temporal information with the
 203 denoising network, denoted as ϵ_θ^v . This network is extended
 204 to 3D by either introducing additional temporal attention
 205 modules [5, 73], or replacing the spatial attention modules
 206 with spatio-temporal ones [75, 86].

207 3.2. Video Understanding Probing Framework

208 Figure 2 illustrates our unified probing framework. We ex-
 209 tract video representations from the denoising module and
 210 subsequently apply a lightweight task-specific head for var-
 211 ious tasks.

212 3.2.1 Diffusion Features

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

$$221 \mathbf{z}_{\text{feature}} = \epsilon_\theta^{v(n)}(\mathbf{z}_T, T), \quad (5)$$

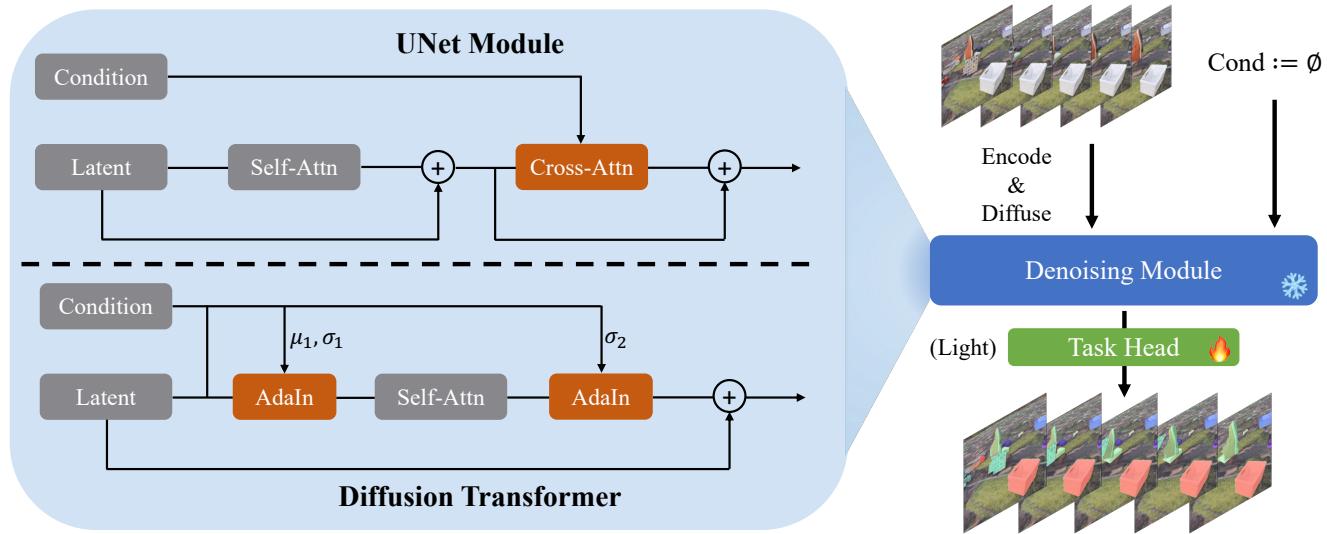


Figure 2. The architecture of our probing framework for video understanding using diffusion models. Video feature representations are extracted from the denoising module, followed by a lightweight task head to produce task-specific annotations. The process of feature extraction from UNet or DiT models (SD3 [15]) is illustrated on the left. Notice that we ignore the timestep input for simplification.

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 use 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. We detail the specific task heads for our evaluated tasks below.

Action Recognition is the task that aims to predict an action label for a given video. Following previous practice [4, 71], we take the averaged feature map and apply a two-layer MLP, where the hidden dimension is the same as the input features, to predict the final label.

Object Discovery identifies and tracks dynamic objects from videos in a self-supervised manner. We adopt the architecture from MoTok [2] where cross-attention layers with learnable queries, called slots [43], are trained to group foreground regions in video with feature-level reconstruction as the learning signal.

Scene Understanding aims to predict pixel-wise scene

Model	Type	Architecture	Dataset	Feature Dim	Downsample
DINOv2 [52]	Image	ViT-L	LVD-142M	1024	14
VideoMAE [71]	Video	ViT-L	Kinetics400-240k	1024	16
VJEPa [4]	Video	ViT-L	VideoMix2M	1024	16
SD [60]	Image	UNet	LAION-5B	1280/640	8/16
SD3 [15]	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
Open-Sora Zheng et al. [88]	Video	DiT	Mix-210M	288	8

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

properties, e.g. semantic labels and depth values, for the given video. Following DINOv2 [52], we directly apply a two-layer MLP on top of the feature map and interpolate them to the original resolution to predict the labels.

Label Propagation is a training-free task where instance masks or keypoints from an initial frame are propagated to each subsequent frame in a video. Rather than predicting new labels, label propagation transfers the initial labels frame-by-frame, leveraging the continuity of appearance across frames. As in prior methods [32, 70], we achieve this by using a k-nearest neighbors (k-NN) search across a feature queue containing the initial frame and the most recent m frames, thus no specialized task head is required.

4. Experimental Evaluations

4.1. Evaluation Settings

Baseline Models. We perform our video understanding analysis with seven visual foundation models. **DINOv2** [52] is a contrastive learning-based image-level foundation model. **VJEPa** [4] and **VideoMAE** [71] learn com-

250
251
252
253
254
255
256
257
258
259
260
261
262

263
264
265
266
267
268

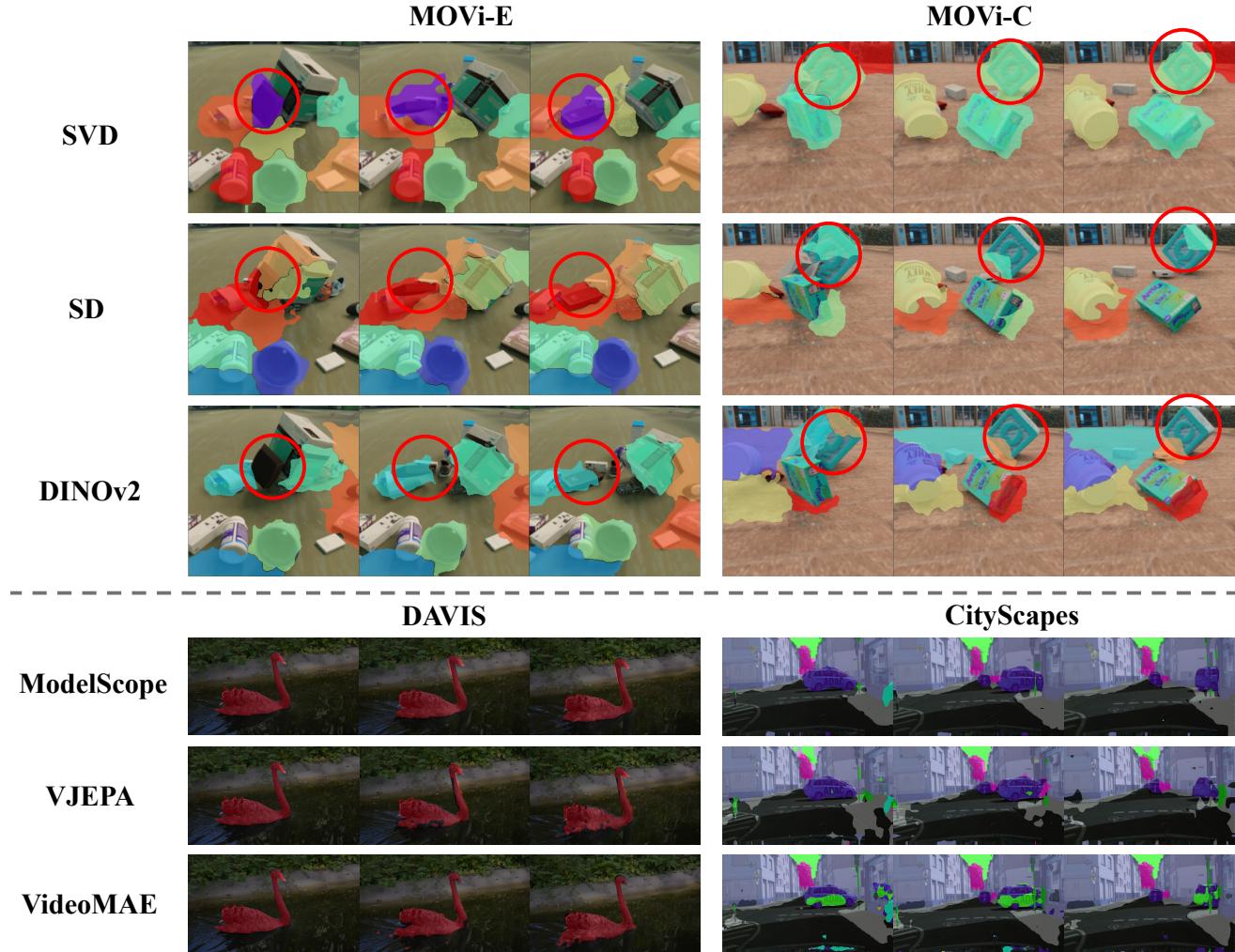


Figure 3. Representative visual comparisons between the results of video diffusion models and other foundation models. **Top:** Video diffusion models capture motion dynamics more effectively than image-based models; **Bottom:** Video diffusion models demonstrate a stronger understanding of world structure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

269 prehensive video representations by reconstructing from
 270 masked video patches. **Stable Diffusion (SD)**[60] and
 271 **Stable Diffusion 3 (SD3)**[15] are text-to-image diffusion
 272 model with UNet [61] and DiT [54] as denoising back-
 273 bones. **ModelScopeT2V** [75] and **Stable Video Diffusion**
 274 (**SVD**) are video diffusion models that take SD as the initial-
 275 ization and further fine-tune on large-scale video data. Addi-
 276 tionally, we include the DiT-based video diffusion model,
 277 Open-Sora [88], in the action recognition evaluation but ex-
 278 clude it from other tasks due to its inability to produce pre-
 279 cise patch-wise representations. Detailed configurations of
 280 these feature extractors are provided in Table 1.

281 **Datasets and Metrics.** We evaluate *action recognition*
 282 recognition with top 1 and top 5 accuracy on UCF101 [69]
 283 and HMDB51 [38]. We study the object discovery task

284 on MOVi-C and MOVi-E [18], and take foreground ad-
 285 just random index (FG. ARI) and video mean best over-
 286 lap (mBO) as metrics. We evaluate the scene under-
 287 standing task with semantic segmentation and depth estima-
 288 tion on CityScapes [13], and take mean interaction over unions
 289 (mIoU) and mean L_2 error (mErr), for the two tasks respec-
 290 tively. We conduct the label propagation for video object
 291 segmentation on DAVIS17 [57] and keypoint estimation on
 292 JHMDB [34] following the same setup as DIFT [70]. We re-
 293 port region-based similarity \mathcal{J} and contour-based accuracy
 294 \mathcal{F} [55] for DAVIS17, and percentage of correct keypoints
 295 (PCK) for JHMDB.

296 **Key Implementation Details.** We use the noise level 50
 297 by default, with a corresponding timestep T=50 (for SD,
 298 ModelScope, and SVD) or T=16 (for SD3 and Open-Sora).

Backbone	UCF101		HMDB51		MOVi-C		MOVi-E		CityScape		DAVIS17			JHMDB	
	Top1 Acc	Top 5 Acc	Top1 Acc	Top 5 Acc	FG.ARI	mBO	FG.ARI	mBO	mIoU(SS)	mErr(DE)	\mathcal{J}_m	\mathcal{F}_m	$\mathcal{J} \& \mathcal{F}_m$	PCK@0.1	PCK@0.2
DINOv2	89.8	97.8	61.6	89.6	55.6	29.2	71.9	26.3	53.6	4.30	64.8	69.1	67.0	50.42	78.71
VideoMAE	87.9	97.9	55.4	83.4	24.5	14.3	32.7	14.1	37.8	5.73	30.5	37.5	34.0	32.51	59.30
VJEPa	92.1	98.5	66.5	92.3	31.8	18.6	49.9	18.0	41.3	5.27	52.3	58.0	55.1	37.55	70.31
SD	63.5	86.1	33.0	68.1	40.6	24.8	63.4	26.9	44.5	4.97	67.8	74.6	71.2	60.48	80.77
SD3	60.9	85.8	32.4	62.1	43.3	26.3	65.1	28.6	46.0	5.09	48.5	54.8	51.6	38.17	65.89
ModelScope	80.6	94.9	50.7	80.2	41.3	25.1	63.7	27.5	49.3	3.98	65.3	72.4	68.4	60.90	82.83
SVD	92.3	98.6	63.8	89.7	44.2	26.7	65.4	29.4	48.1	4.68	59.8	67.7	63.8	60.52	81.84
Open-Sora	47.3	75.9	22.1	54.8	-	-	-	-	-	-	-	-	-	-	-

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.

For the layer index, we design the use of block index 1 (for SD, ModelScope, and SVD) and layer index 12 (for SD3 and Open-Sora) for action recognition. For the other tasks, we use block index 2 and layer index 24 respectively. 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 ModelScope to fit its CUDA requirement.

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

4.2. Main Results

We show the quantitative results for our main evaluation of the four tasks in Table 2, and representative visual comparisons in Figure 3. More visualizations are included in the supplementary material.

Comparisons between ModelScope and SVD. For the following discussions, we treat ModelScope and SVD as variants of the same “video diffusion model” category, despite differences in their model type (Text-to-Video vs. Image-to-Video), for which we use unconditional versions to minimize conditioning effects. While their performance varies across tasks – likely due to differences in training data and fine-tuning strategies (ModelScope fine-tunes only temporal modules, while SVD uses full fine-tuning) – these variations make it challenging to draw universal conclusions based on specific tasks. Given the lack of a standardized training strategy, we focus on their shared foundations instead: both models are based on SD with additional video training, which enables us to discuss their common strengths and limitations in video understanding.

Overall Conclusions. Across all four tasks, video diffusion models consistently rank among the top performers, highlighting their robustness and adaptability in video understanding. As illustrated by the visual comparisons in Figure 3, video diffusion models capture motion dynamics more effectively than image-based models and 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.

Action Recognition. Surprisingly, SVD achieves the highest performance on UCF101 and ranks second on HMDB51, consistently outperforming both image diffusion models and the conventional DINOv2 and VideoMAE encoders. This result highlights the ability of well-trained video diffusion models to capture global-level video representations effectively. However, Open-Sora and SD3, which use DiT architectures, exhibit suboptimal performance. A potential reason may lie in how DiT models fuse multi-modal features, suggesting an open research challenge: developing improved feature extraction techniques tailored for DiT-based diffusion models.

Object Discovery. Overall DINOv2 achieves the highest performance among all models, demonstrating its superior object-awareness. However, it is worth noticing that video diffusion models outperform in terms of mBO on the MOVi-E dataset which involves more complex ego and object motion. This suggests that diffusion models are particularly effective at identifying and tracking objects in challenging motion scenarios, making them especially useful for tasks requiring precise localization and tracking. Visual comparisons in Figure 3 provide further evidence where SVD precisely tracks objects with complicated motion.

Scene Understanding. ModelScope and DINOv2 emerge as the top performers in these tasks, with DINOv2 excelling in semantic understanding and ModelScope showing superior performance in depth estimation. For ModelScope, we hypothesize that its success stems from its ability to leverage motion information, which inherently aids in understanding depth.

Label Propagation. On DAVIS17, video diffusion models generally lag behind their image-based counterparts. 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, a key challenge in video object segmentation (VOS). In contrast, on the JHMDB dataset, where pose estimation focuses solely on a single moving object, video

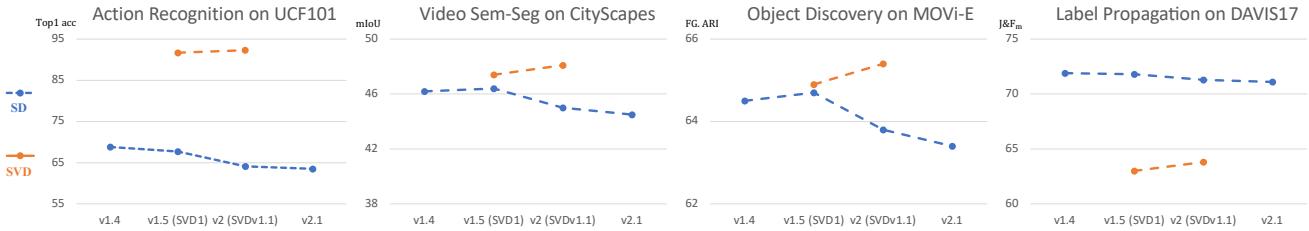


Figure 4. Comparison between generation ability and downstream task performance on SD and SVD series. The later SVD checkpoint consistently improves performance across all tasks while the 1-series SD models generally outperform the 2-series models. These results indicate that greater generative capacity does not necessarily translate to improved performance in visual perception tasks.

378 diffusion models demonstrate their strengths.

379 In summary, video diffusion models provide semantic-
380 and geometric-aware representations that contain both high-
381 level abstractions and detailed information, positioning
382 them as unique and competitive candidates for video un-
383 derstanding.

384 4.3. Guidelines for Video Diffusion Adoption

385 4.3.1 Optimal Use of Video Diffusion Models

386 In our main evaluation, we use frozen video diffusion re-
387 presentations with fixed noise levels and layer indices. In this
388 section, we investigate how to better adapt these represen-
389 tations for video understanding by providing guidance on
390 layer selection and fine-tuning strategies.

391 **Noise Levels and Block Indices.** We examine the ef-
392 ffects of noise level selection and block indices in SVD for
393 action recognition on HMDB51 and label propagation on
394 DAVIS17, as summarized in Table 3. The results suggest
395 that noise level plays a relatively smaller and task-specific
396 role compared to block indices. Generally, a small amount
397 of noise (*e.g.*, corresponding to $T = 50$) yields strong
398 results. In contrast, block indices significantly influence
399 downstream task performance: features from earlier blocks
400 encode abstract, high-level information, making them ideal
401 for classification tasks, while features from later blocks cap-
402 ture finer details, benefiting dense prediction tasks. These
403 findings are consistent with observations from image diffu-
404 sion models, as reported by Tang et al. [70].

405 **Fine-Tuning Video Diffusion Models.** For certain video
406 understanding tasks, fine-tuning the backbone is essential
407 and typically results in improved performance. To ex-
408 plore the impact and strategies of fine-tuning video diffu-
409 sion models for perception tasks, we fine-tune the SVD de-
410 noising UNet on HMDB51 and MOVi-E. The results are
411 summarized in Table 4 (first two rows). Notably, for object
412 discovery, we slightly modify the baseline architecture [2],
413 with details provided in the supplementary material. As a
414 result, the reported FG.ARI score for the frozen model dif-
415 fers from that in Table 2.

Noise Level	Block Index	HMDB51		DAVIS17		
		Top1 Acc	Top5 Acc	\mathcal{J}_m	\mathcal{F}_m	$\mathcal{J} \& \mathcal{F}_m$
0	1	60.3	88.0	52.1	44.9	48.5
50	1	63.8	89.7	51.1	42.6	46.9
100	1	63.9	89.4	50.3	41.6	46.0
200	1	62.6	88.7	50.2	41.3	45.8
0	2	31.1	64.0	60.8	68.0	64.4
50	2	33.7	66.9	59.8	67.7	63.8
100	2	35.4	68.0	59.6	67.2	63.4
200	2	32.8	66.8	59.1	64.5	62.8

Table 3. Ablation on noise level selection and block index of SVD on HMDB51 and DAVIS17. 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.

416 Notably, by comparing the change of parameters of all
417 the modules, we find that the last in-use upsampling block
418 (*i.e.* block index 1 for action recognition and block index 2
419 for object discovery) exhibits the highest sensitivity to pa-
420 rameter changes, highlighting their critical role in enhanc-
421 ing task performance. Inspired by previous efficient diffu-
422 sion fine-tuning approaches [3, 39, 62], we construct two
423 fine-tuning variants: one incorporates LoRA [30] adapta-
424 tion layers in all cross-attention blocks, while the other fine-
425 tune only the most sensitive upsampling block. The results
426 for these two variants are reported in Table 4 (last two rows).
427 These findings demonstrate that efficient fine-tuning strate-
428 gies can significantly enhance performance while keeping
429 training costs reasonable, offering practical guidance for
430 optimizing video diffusion models.

431 4.3.2 Generation V.S. Perception

432 In this section, we explore an intriguing question: *does a*
433 *diffusion model with superior generation capacity inher-*
434 *ently perform better in visual perception tasks?* While
435 we could evaluate the generative capacity of different dif-
436 fusion models directly, this approach is challenging due
437 to their diverse conditioning mechanisms –some are text-

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 ×
Sensitive	67.1	1.3 ×	1.8 ×	68.1	1.4 ×	1.9 ×

Table 4. Performance and training cost for finetuning SVD UNet. “Sensitive” denotes only fine-tuning the most sensitive UNet block (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.

438 conditioned, others image-conditioned – and their application
439 across both image and video generation. Instead, we
440 adopt an alternative strategy: comparing the performance of
441 different checkpoints of the same model, under the assumption
442 that later versions exhibit improved generative capacity.
443

We use SD and SVD as backbone models and evaluate four versions of SD (v1.4, 1.5, 2.0, and 2.1) alongside two versions of SVD (v1 and v1.1) across the four tasks, with results summarized in Figure 4. For SVD, the later checkpoint consistently improves performance across all tasks, aligning with its enhanced generative capacity. However, for SD, the 1-series models generally outperform the 2-series models, though the optimal version varies by task. This discrepancy may stem from differences in the scale and composition of training data across versions.

Overall, these results suggest that greater generative capacity does not necessarily translate to improved performance in visual perception tasks, indicating that there is no universal metric for selecting a representation exists as of yet.

458 4.4. Discussions

459 **Inference Cost.** We report the inference time and memory
460 usage for a single batch of size [6, 256, 256] on the MOVi-
461 E dataset, using an NVIDIA A100 GPU in Table 5. The
462 baseline model, DINOv2, has an inference time of 0.224
463 seconds and consumes 2.6 GB of GPU memory. Notably,
464 the memory consumption for ModelScope is an outlier, due
465 to the lack of optimization in its public implementation.
466 In general, diffusion-based and video-based models require
467 more computational resources, though these costs remain
468 acceptable. The exception is SD3, which employs a DiT-
469 based architecture. This observation is consistent with our
470 earlier conclusions and highlights the need to develop more
471 efficient and effective feature extraction methods for DiT-
472 based models.

473 **Limitations of Video Diffusion Representations.** We
474 show two typical limitations for video diffusion representa-
475 tions on label propagation in Figure 5: difficulty in handling

Model	DINOv2	VideoMAE	VJEPAs	SD	SD3	ModelScope	SVD
Memory	1.0×	1.7 ×	1.1×	1.8×	4.6×	8.3×	2.7×
Inf. time	1.0×	2.1 ×	1.7×	1.1×	3.3×	2.0×	2.1×

Table 5. Time and Memory Consumptions for all the compared models. Results are tested on the MOVi-E dataset with a single batch with dimensions [6, 256, 256]. Diffusion-based and video-based models require more computational resources but the costs remain acceptable.



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.

476 occlusion among instances of the same semantic category,
477 and challenges with distinguishing nearby objects that share
478 similar motion.

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.

Two feasible **future work** of this study include: (1) designing a more advanced feature extraction pipeline with newly introduced DiT-based models. (2) Exploring other ways of leveraging video diffusion models beyond merely using them as encoders [1, 79].

Social Impact. By pushing the boundaries of what is possible with video diffusion models, the findings in this paper can further inspire future explorations with video diffusion models in both generative and video analysis aspects.

503

References

504

- [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. 1, 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, 3, 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] 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. 2
- [9] 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
- [10] 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
- [11] 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
- [12] 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
- [13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5
- [14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 2
- [15] 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, 5
- [16] 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
- [17] 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
- [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. 3
- [26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

- 617 [27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion
618 probabilistic models. *NeurIPS*, 2020. 2, 3
- 619 [28] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang,
620 Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben
621 Poole, Mohammad Norouzi, David J Fleet, et al. Imagen
622 video: High definition video generation with diffusion
623 models. *arXiv preprint arXiv:2210.02303*, 2022. 2
- 624 [29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William
625 Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion
626 models. *arXiv preprint arXiv:2204.03458*, 2022. 2
- 627 [30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-
628 Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.
629 Lora: Low-rank adaptation of large language models. *arXiv
630 preprint arXiv:2106.09685*, 2021. 7
- 631 [31] Xun Huang and Serge Belongie. Arbitrary style transfer in
632 real-time with adaptive instance normalization. In *ICCV*,
633 2017. 3
- 634 [32] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time
635 correspondence as a contrastive random walk. *NeurIPS*,
636 2020. 4
- 637 [33] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat
638 Behl. Peekaboo: Interactive video generation via masked-
639 diffusion. *arXiv preprint arXiv:2312.07509*, 2023. 2
- 640 [34] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black.
641 Towards understanding action recognition. In *ICCV*, 2013. 5
- 642 [35] Levon Khachatryan, Andranik Movsisyan, Vahram Tade-
643 vosyan, Roberto Henschel, Zhangyang Wang, Shant
644 Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-
645 image diffusion models are zero-shot video generators. *arXiv
646 preprint arXiv:2303.13439*, 2023. 2
- 647 [36] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi,
648 Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Seg-
649 ment like me. *arXiv preprint arXiv:2309.03179*, 2023. 3
- 650 [37] Diederik P Kingma and Max Welling. Auto-encoding vari-
651 ational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2
- 652 [38] Hildegard Kuehne, Hueihan Jhuang, Estibaliz Garrote,
653 Tomaso Poggio, and Thomas Serre. Hmdb: a large video
654 database for human motion recognition. In *ICCV*, 2011. 5
- 655 [39] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli
656 Shechtman, and Jun-Yan Zhu. Multi-concept customization
657 of text-to-image diffusion. In *CVPR*, 2023. 7
- 658 [40] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 2
- 659 [41] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis
660 Brown, and Deepak Pathak. Your diffusion model is secretly
661 a zero-shot classifier. In *ICCV*, 2023. 1
- 662 [42] Ruoshi Liu, Rundi We, Basile Van Hoorick, Pavel Tok-
663 makov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3:
664 Zero-shot one image to 3d object. In *ICCV*, 2023. 3
- 665 [43] Francesco Locatello, Dirk Weissenborn, Thomas Unter-
666 thiner, Aravindh Mahendran, Georg Heigold, Jakob
667 Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-
668 centric learning with slot attention. *NeurIPS*, 2020. 4
- 669 [44] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holyn-
670 ski, and Trevor Darrell. Diffusion hyperfeatures: Search-
671 ing through time and space for semantic correspondence.
672 *NeurIPS*, 2023. 3
- 673 [45] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang,
674 Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and
675 Tieniu Tan. Videofusion: Decomposed diffusion mod-
676 els for high-quality video generation. In *Proceedings of
677 the IEEE/CVF Conference on Computer Vision and Pattern
678 Recognition*, pages 10209–10218, 2023. 2
- 679 [46] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert,
680 Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing
681 visual foundation models for complex 3d scene under-
682 standing. *arXiv preprint arXiv:2409.03757*, 2024. 3
- 683 [47] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jia-
684 jun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided
685 image synthesis and editing with stochastic differential equa-
686 tions. *arXiv preprint arXiv:2108.01073*, 2021. 3
- 687 [48] Sauradip Nag, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song,
688 and Tao Xiang. Diffad: Temporal action detection with pro-
689 posal denoising diffusion. *arXiv preprint arXiv:2303.14863*,
690 2023. 3
- 691 [49] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and
692 Seung Wook Kim. Emerdiff: Emerging pixel-level seman-
693 tic knowledge in diffusion models. *arXiv preprint
694 arXiv:2401.11739*, 2024. 3
- 695 [50] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and
696 Martin Renqiang Min. Conditional image-to-video genera-
697 tion with latent flow diffusion models. In *CVPR*, 2023. 2
- 698 [51] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav
699 Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and
700 Mark Chen. Glide: Towards photorealistic image generation
701 and editing with text-guided diffusion models. *arXiv preprint
702 arXiv:2112.10741*, 2021. 2
- 703 [52] Maxime Oquab, Timothée Darcret, Théo Moutakanni, Huy
704 Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez,
705 Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al.
706 Dinov2: Learning robust visual features without supervision.
707 *TMLR*, 2023. 1, 4
- 708 [53] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal
709 Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt:
710 Amodal segmentation by synthesizing wholes. In *CVPR*,
711 2024. 3
- 712 [54] William Peebles and Saining Xie. Scalable diffusion models
713 with transformers. In *ICCV*, 2023. 1, 2, 3, 5
- 714 [55] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc
715 Van Gool, Markus Gross, and Alexander Sorkine-Hornung.
716 A benchmark dataset and evaluation methodology for video
717 object segmentation. In *CVPR*, 2016. 5
- 718 [56] Pietro Perona and Jitendra Malik. Scale-space and edge de-
719tection using anisotropic diffusion. *PAMI*, 1990. 2
- 720 [57] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Ar-
721 beláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017
722 davis challenge on video object segmentation. *arXiv preprint
723 arXiv:1704.00675*, 2017. 2, 5
- 724 [58] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie
725 Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hier-
726 archical spatio-temporal decoupling for text-to-video gener-
727 ation. *arXiv preprint arXiv:2312.04483*, 2023. 2
- 728 [59] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu,
729 and Mark Chen. Hierarchical text-conditional image gen-

- 730 eration with clip latents. *arXiv preprint arXiv:2204.06125*,
731 2022. 2
- 732 [60] Robin Rombach, Andreas Blattmann, Dominik Lorenz,
733 Patrick Esser, and Björn Ommer. High-resolution image syn-
734 thesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3,
735 4, 5
- 736 [61] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net:
737 Convolutional networks for biomedical image segmentation.
738 In *MICCAI*, 2015. 3, 5
- 739 [62] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch,
740 Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine
741 tuning text-to-image diffusion models for subject-driven
742 generation. In *CVPR*, 2023. 7
- 743 [63] Chitwan Saharia, William Chan, Saurabh Saxena, Lala
744 Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed
745 Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi,
746 Rapha Gontijo Lopes, et al. Photorealistic text-to-image
747 diffusion models with deep language understanding. *arXiv*
748 preprint *arXiv:2205.11487*, 2022. 2
- 749 [64] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek
750 Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet.
751 The surprising effectiveness of diffusion models for optical
752 flow and monocular depth estimation. In *NeurIPS*, 2023. 3
- 753 [65] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and
754 David J Fleet. Monocular depth estimation using diffusion
755 models. *arXiv preprint arXiv:2302.14816*, 2023. 3
- 756 [66] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An,
757 Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual,
758 Oran Gafni, et al. Make-a-video: Text-to-video generation
759 without text-video data. *arXiv preprint arXiv:2209.14792*,
760 2022. 2
- 761 [67] Jiaming Song, Chenlin Meng, and Stefano Ermon.
762 Denoising diffusion implicit models. *arXiv preprint*
763 *arXiv:2010.02502*, 2020. 2
- 764 [68] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Ab-
765 hishek Kumar, Stefano Ermon, and Ben Poole. Score-based
766 generative modeling through stochastic differential equa-
767 tions. *arXiv preprint arXiv:2011.13456*, 2020. 2
- 768 [69] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah.
769 Ucf101: A dataset of 101 human actions classes from videos
770 in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5
- 771 [70] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng
772 Phoo, and Bharath Hariharan. Emergent correspondence
773 from image diffusion. *NeurIPS*, 2023. 3, 4, 5, 7
- 774 [71] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang.
775 Videomae: Masked autoencoders are data-efficient learners
776 for self-supervised video pre-training. In *NeurIPS*, 2022. 1,
777 4
- 778 [72] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan
779 Salakhutdinov. Effective data augmentation with diffusion
780 models. In *ICLR*, 2024. 3
- 781 [73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszk-
782 reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia
783 Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3
- 784 [74] Cong Wang, Jiaxi Gu, Panwen Hu, Songcen Xu, Hang Xu,
785 and Xiaodan Liang. Dreamvideo: High-fidelity image-to-
786 video generation with image retention and text guidance.
787 *arXiv preprint arXiv:2312.03018*, 2023. 2
- 75 [75] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang,
788 Xiang Wang, and Shiwei Zhang. Modelscope text-to-video
789 technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1,
790 2, 3, 4, 5
- 791 [76] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong
792 Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic
793 image synthesis via diffusion models. *arXiv preprint*
794 *arXiv:2207.00050*, 2022. 2
- 795 [77] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing,
796 Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and
797 Nong Sang. A recipe for scaling up text-to-video genera-
798 tion with text-free videos. *arXiv preprint arXiv:2312.15770*,
799 2023. 2
- 800 [78] Joachim Weickert et al. *Anisotropic diffusion in image pro-
801 cessing*. Teubner Stuttgart, 1998. 2
- 802 [79] Zejia Weng, Xitong Yang, Zhen Xing, Zuxuan Wu, and
803 Yu-Gang Jiang. Genrec: Unifying video generation
804 and recognition with diffusion models. *arXiv preprint*
805 *arXiv:2408.15241*, 2024. 3, 8
- 806 [80] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian
807 Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu
808 Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning
809 of image diffusion models for text-to-video generation. In
810 *ICCV*, 2023. 2
- 811 [81] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiao-
812 long Wang, and Shalini De Mello. ODISE: Open-vocabulary
813 panoptic segmentation with text-to-image diffusion models.
814 In *CVPR*, 2023. 1, 3
- 815 [82] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gun-
816 jan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yin-
817 fei Yang, Burcu Karagol Ayan, et al. Scaling autoregres-
818 sive models for content-rich text-to-image generation. *arXiv*
819 preprint *arXiv:2206.10789*, 2022. 2
- 820 [83] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei,
821 Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel
822 Albanie, and Dong Ni. Instructvideo: Instructing video
823 diffusion models with human feedback. *arXiv preprint*
824 *arXiv:2312.12490*, 2023. 2
- 825 [84] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Pola-
826 nia Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan
827 Yang. A tale of two features: Stable diffusion complements
828 dino for zero-shot semantic correspondence. *NeurIPS*, 2023.
829 3
- 830 [85] Lvmin Zhang and Maneesh Agrawala. Adding conditional
831 control to text-to-image diffusion models. *arXiv preprint*
832 *arXiv:2302.05543*, 2023. 2
- 833 [86] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao,
834 Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and
835 Jingren Zhou. I2vgen-xl: High-quality image-to-video
836 synthesis via cascaded diffusion models. *arXiv preprint*
837 *arXiv:2311.04145*, 2023. 2, 3
- 838 [87] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie
839 Zhou, and Jiwen Lu. Unleashing text-to-image diffusion
840 models for visual perception. In *ICCV*, 2023. 1, 3
- 841 [88] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen,
842 Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang
843 You. Open-sora: Democratizing efficient video production
844 for all, 2024. 1, 2, 4, 5
- 845

- 846 [89] Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan,
847 Chunming Qiao, and Gang Hua. Exploring pre-trained text-
848 to-video diffusion models for referring video object segmen-
849 tation. *arXiv preprint arXiv:2403.12042*, 2024. 3