THE REPRESENTATIONAL POWER OF FRONTIER VIDEO DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Do video diffusion models learn useful representations? Prior work suggests that diffusion representations capture low-level geometry but struggle with high-level semantics. We demonstrate that frontier video diffusion models overcome this limitation. By systematically probing their intermediate activations using zero-shot and parametric probes, we reveal that visual representations evolve across both network depth and noise levels, forming a distributed hierarchical feature space. Importantly, these representations are easily accessible with a single forward pass, yielding state-of-the-art or competitive performance on semantic and geometric understanding tasks without any fine-tuning. Our findings pave the way towards unified backbones for both video generation and perception.

1 INTRODUCTION

Current paradigms in visual representation learning struggle to reconcile geometry with semantics. Contrastive methods and large Vision-Language Models (VLMs) (Radford et al., 2021; Beyer et al., 2024; Gemini Team, 2025) operate primarily in language space, yielding strong semantics but neglecting precise spatio-temporal details. Conversely, pixel-level reconstruction approaches (Tong et al., 2022; Wang et al., 2023; Carreira et al., 2025; Zhang et al., 2026) excel at preserving local geometry and motion but cannot easily support broad semantic tasks. A single, task-agnostic foundation video model that supports both types of capabilities remains an open problem.

We investigate the representational power of frontier video generative transformers as a candidate for a general-purpose video model. We argue that generating temporally and geometrically consistent video requires the model to implicitly learn the underlying mechanics of the physical world, such as motion, object permanence, and part-object relationships. However, recent studies probing frozen video diffusion representations concluded that they capture low-level geometry but struggle with high-level semantics, casting doubt on their eligibility as general-purpose perception encoders (Vélez et al., 2025; Zhu et al., 2026).

In this work, we demonstrate that state-of-the-art video diffusion models do, in fact, overcome this limitation as they become more capable at generating physically and temporally consistent videos. By systematically investigating the intermediate activations of both proprietary models like Veo3 (Google DeepMind, 2025) and open-weight models like Wan2.2-T2V-A14B (Team Wan et al., 2025), we map the hierarchy of their latent spaces using zero-shot metrics (Huh et al., 2024; Zhu et al., 2026) and lightweight trainable probes. We identify optimal spots for extracting representations with a single forward pass through the diffusion backbone. This makes it as computationally efficient as a standard discriminative visual encoder, e.g. (Carreira et al., 2025; Wang et al., 2024b), circumventing the high inference costs typically associated with iterative denoising. To emphasise that these representations are extracted via a single forward pass without iterative denoising, we denote these models as *Veo3-1P* and *Wan2.2-1P* where *1P* stands for one-pass. By pairing these frozen representations with lightweight decoders, we can achieve strong perception performance without altering the pre-trained generative weights. Our extensive evaluations demonstrate that frozen video diffusion models can function as general-purpose video encoders, paving the way towards unified backbones for both video generation and understanding.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 RELATED WORK

The most common paradigms for learning visual representations are variations of masked autoencoding (MAE) (He et al., 2021; Tong et al., 2022; Wang et al., 2023), scaled in 4DS (Carreira et al., 2025) and D4RT (Zhang et al., 2026), and contrastive learning, with or without negative pairs (SigLIP (Zhai et al., 2023), SimCLR (Chen et al., 2020), CLIP (Radford et al., 2021), BYOL (Grill et al., 2020), BRAVE (Recasens et al., 2021), DINO (Caron et al., 2021) to name a few). These methods generally strike a trade-off between low level scene understanding and high level semantics, failing to support both. Even hybrid approaches (Zhao et al., 2024; Wang et al., 2024b; Papalampidi et al., 2023) that combine masked auto-encoding (MAE) with contrastive learning can still accurately span only one category of tasks. More recently, the V-JEPA line of works (Bardes et al., 2024; Assran et al., 2025; Mur-Labadia et al., 2026) focus on predicting the future in latent space to avoid the computational cost of pixel-level prediction, and show strong results on a wide range of understanding tasks. These pre-training methods lead to strong results on downstream tasks, but most of the time they require extensive fine-tuning or have fairly complex pre-training pipelines, balancing different optimisation objectives.

Driven by the impressive realism and diversity of the outputs generated by diffusion models (Rombach et al., 2022; Gupta et al., 2023; Google DeepMind, 2025; Team Wan et al., 2025), multiple works have started to investigate their understanding capabilities. In (Li et al., 2023a) the authors show that a diffusion-based setup can perform zero-shot classification by conditioning the generative model with the possible classes and selecting the class that leads to the best reconstruction of the noise added to the input image. In (Luo et al., 2023), the authors build diffusion hyper-features by combining the activations from different blocks and noise levels to build robust pixel-level descriptors. In (Wiedemer et al., 2025), the authors show impressive generation results by prompting the Veo3 model with challenging tasks that require geometry and physics understanding, hinting that Veo3 has implicitly learnt a world model. In addition, works like (Ha & Schmidhuber, 2018), (Hafner et al., 2025), (Bruce et al., 2024; Parker-Holder et al., 2024) confirm that generative models support learning a world model for simulated environments, with representations that can be used to inform action policies.

The closest work to our approach is (Vélez et al., 2025). The authors take variants of the Walt model (Gupta et al., 2023) adapted for image and video, and repurpose their activations for visual understanding tasks. They conclude that diffusion models excel at lower level understanding problems like depth or camera pose estimation, but struggle with high-level semantic tasks. Similarly, in (Zhu et al., 2026), the authors measure the alignment of representations produced by different visual encoders amongst themselves and with language embeddings. Interestingly, their findings suggest that diffusion models like Walt produce representations that align with established visual encoders like DINOv2 (Oquab et al., 2024) and VideoMAEv2 (Wang et al., 2023), but the alignment with text embeddings is significantly lower. In our work, we show that the recent generation of video diffusion models have strong semantic awareness, in addition to geometry and motion understanding, constituting a strong foundation for a complete perception system. We show state-of-the-art results on video classification on the challenging SSv2 dataset (Goyal et al., 2017), compared to strong baselines from the other families of visual pre-training methods (MAE, contrastive, latent prediction). To further the investigation of semantic understanding, we set up captioning experiments, showing satisfactory performance compared to a strong SigLIP-based baseline (Zhai et al., 2023) from the PaliGemma family (Beyer et al., 2024).

The interplay between generation and understanding started to receive more attention recently. For example, detailed point tracking conditioning leads to more consistent generation (Jeong et al., 2025), or detailed language captioning produced by Gemini models lead to improved Veo generation (Google DeepMind, 2025). Transfusion (Zhou et al., 2025) elegantly combines generation and understanding pre-training into a single model, allowing the two objectives to support each other. However, in all these works, the main finding has been that advanced understanding models lead to significant improvement in generative models. The other direction, how can generation help understanding, has so far been explored mainly for low-level geometry tasks. In our work, we show that frozen diffusion generative models can power video understanding and achieve competitive performance on a wide range of visual tasks, while being able to generate high-quality videos. The implications could be substantial from a practical point of view, as it could lead to one visual model shared across generation and understanding. We estimate that this finding can be very impactful for the community, by encouraging more research into understanding the representation power of diffusion models.

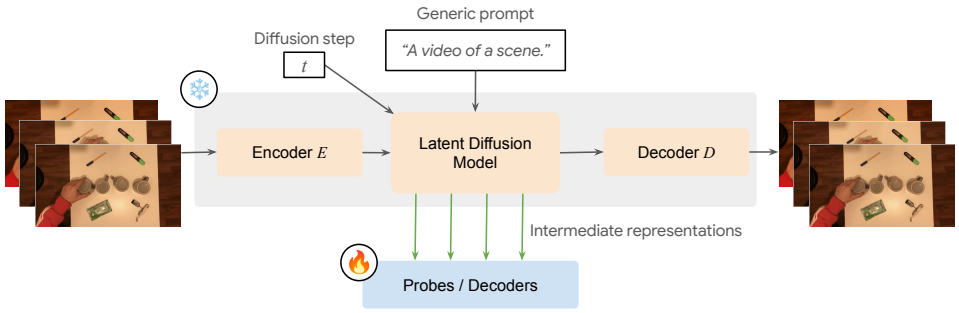


Figure 1: Video generative model repurposed as video encoder.

3 THE STRUCTURE OF THE GENERATIVE MANIFOLD

The iterative diffusion generation requires many denoising steps to generate a video. This slow sequential operation mode is not suitable for video encoding. To design efficient video feature extractors based on video diffusion models, we conduct a detailed investigation of the structure of the latent embeddings spawned by these models, in search for the optimal depth and noise level from where to extract representations with a single forward pass through the model. After a brief overview of diffusion models, we detail the zero-shot and trained light probes used for our analysis, and we conclude this section with a summary discussion of the findings.

3.1 BACKGROUND ON LATENT DIFFUSION MODELS

Video Latent Diffusion Models (LDMs) like Veo3 operate within a compressed, lower-dimensional representation space. During training, given an input video $x \in \mathbb{R}^{F\times H\times W\times C}$ (containing F frames), an encoder E maps the data into a spatio-temporal latent code $z_0 = E(x)$. A forward process corrupts this latent representation over T steps. At an arbitrary noise level $t \in [0, T]$, the corrupted latent z_t is generated by adding Gaussian noise to z_0: $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_t$ dictates the variance schedule. To invert this process, a backbone network, typically a Diffusion Transformer (DiT) or a U-Net, denoted as f_θ, is trained to predict the injected noise ϵ. The network is conditioned on the corrupted latent z_t, the noise step t, and contextual information c (e.g., text embeddings). The model is optimised via the standard re-weighted variational lower bound: $\mathcal{L}_{LDM} = \mathbb{E}_{z_0,\epsilon,t}\left[||\epsilon - f_\theta(z_t, t, c)||_2^2\right]$. Note that flow matching models like Wan 2.2, also included in our study, replace this stochastic forward process with a continuous-time deterministic framework based on Optimal Transport. In this setup, for a continuous time step $t \in [0, 1]$, the corrupted latent is a linear interpolation between the data and a base noise distribution: $z_t = (1-t)z_0 + t\epsilon$. Rather than predicting the injected noise, the network f_θ is optimised to regress the velocity vector field $f_t = \frac{dz_t}{dt} = \epsilon - z_0$ that transports the data to the noise. The model is trained via the flow matching objective: $\mathcal{L}_{FM} = \mathbb{E}_{z_0,\epsilon,t}\left[||f_\theta(z_t, t, c) - (\epsilon - z_0)||_2^2\right]$.

The backbone f_θ consists of a sequence of L layers (e.g., transformer blocks in the case of a DiT). We denote the intermediate spatio-temporal activations extracted from the l^{th} layer at noise step t as $h_t^{(l)}$. During inference, the model iteratively denoises pure Gaussian noise z_T to recover an estimated clean latent $\hat{z}_0$. Finally, $\hat{z}_0$ is mapped back to pixel space via a decoder D, yielding the generated video $\hat{x} = D(\hat{z}_0)$.

Fig. 1 illustrates our diffusion-based video encoder. We condition the diffusion transformer on a generic text embedding: e.g. *A video of a scene*, to not leak the ground truth (action label or caption) for semantic tasks. We have also experimented with conditioning on an empty text string, without significant difference in results. To deal with the inherent stochasticity of the diffusion process, we use a fixed random seed throughout all our experiments to ensure reproducibility. In all our plots, we report results for blocks and noise levels expressed as percentages obtained as a linear fraction of the total depth or noise steps.

3.2 ZERO-SHOT PROBES

PCA visualisation: To qualitatively assess the evolution of the diffusion features along network depth and noise levels, we plot in Fig. 2 the PCA visualisation of the activations extracted from different depths and noise levels in Veo3-1P. We can observe that at high levels of noise, the features are simple, low-frequency, encoding the general shapes in the scene. At lower noise levels, the features become more complex, encoding higher-frequency, more refined details. This is in line with the intuition that diffusion is spectral autoregression (Dieleman, 2024): at each denoising step, the model learns to produce the features at the next (higher) frequency level.

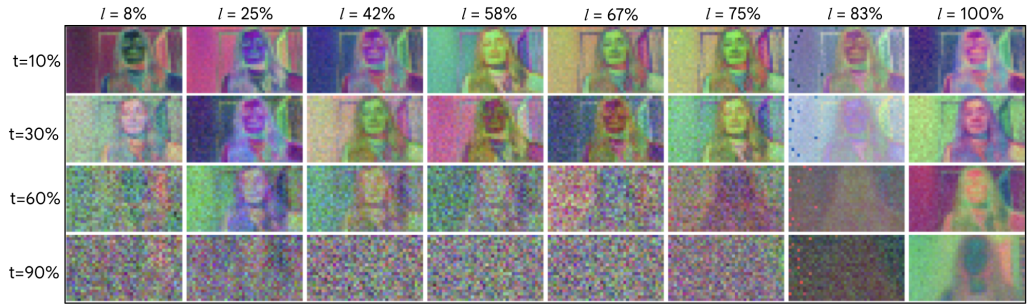


Figure 2: PCA visualisation of activations extracted from Veo3 at different depths and noise levels. For each (depth, noise level) pair, we compute PCA over all spatial tokens aggregated across the dataset. The top three principal components are mapped to RGB channels. Noise level $t = 60\%$ has maximal alignment with language (see Fig. 3 left) and other visual encoders; see text for details.

Shape-texture bias: To quantify this hierarchical structure, we investigate the shape versus texture bias of these representations using the cue conflict dataset (Geirhos et al., 2019). We measure k-nearest neighbour shape-class purity across blocks and noise levels ($k = 10$); see Fig. 3 (left) for Veo3-1P, and Fig. 8 for Wan2.2-1P. We observe that at high noise levels, the representations exhibit a strong shape bias, competitive with DINOv2. This bias decreases with the noise level, mirroring the coarse-to-fine nature of the generative diffusion process: heavy noise destroys high-frequency details, forcing the model to rely on global structural priors (shape) to initiate reconstruction. Then, during lower noise steps, the network resolves local, high-frequency textures. Note that this latent hierarchy organisation is fundamentally different from the hierarchical representations learnt by convolutional neural networks, where features abstraction increases with network depth.

Mutual k-NN alignment: To further assess the structure of the diffusion embedding space, we adopt the zero-shot Mutual k-Nearest Neighbours (MkNN) alignment metric proposed by (Huh et al., 2024) and extended to video in (Zhu et al., 2026). The key idea is to compare the neighbourhood structure induced by the diffusion model's intermediate representations against that of established, independently trained encoders, without requiring any learned mapping between the two spaces.

Concretely, given a dataset of N videos, paired with ground truth captions, we extract a video-level embedding from each video using both the diffusion model and a reference encoder. For the diffusion model, we obtain intermediate activations $h_t^{(l)}$ at a chosen layer l and noise level t, and aggregate them spatio-temporally to produce a single vector per video (see appendix A for details). For the reference encoder, which may operate in a different modality (e.g., a language model) and produce embeddings of a different dimensionality, we similarly obtain one vector per video. In particular, for image models, we follow (Zhu et al., 2026) and average the individual image features across input frames. We then independently construct the k-nearest-neighbour graph in each embedding space and measure the average overlap between the two sets of neighbours. Intuitively, a high overlap indicates that the diffusion model organises videos in a manner consistent with the reference encoder, despite never having been trained with the same objective. More details are included in the appendix A. Following prior work (Huh et al., 2024; Zhu et al., 2026), we additionally optimise over the choice of intermediate layers in both encoders and report the pair of layers that maximises the alignment score. We use k=10 and N=1024 videos.

In Fig. 3, we show the alignment of Veo3-1P and Wan2.2-1P respectively, against the Gemma2-9B-it text encoder (Gemma Team, 2024) on the VATEX dataset. This metric has been shown to correlate

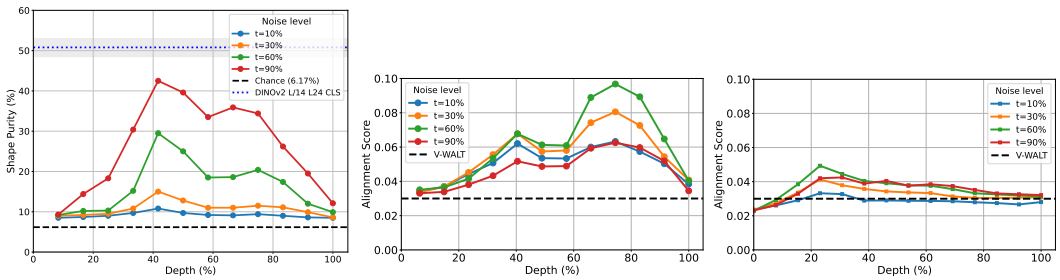
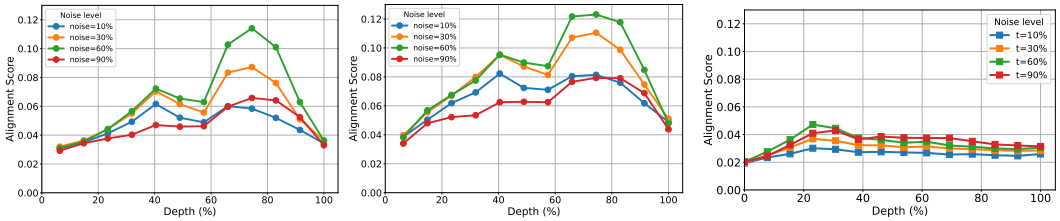


Figure 3: **Left:** Shape-class purity for k-nearest neighbours in shape-texture bias dataset for Veo3-1P compared to DINOv2. **Middle, right:** Mutual-k-NN zero-shot video-text alignment for Veo3-1P and Wan 2.2-1P against Gemma2-9B-it, compared to the best alignment obtained with V-WALT (Vélez et al., 2025) for reference.



Veo3-1P vs DINOv2 | Veo3-1P vs VideoMAEv2 | Wan2.2-1P vs VideoMAEv2

Figure 4: Alignment between diffusion and discriminative representations using mutual k-NN metric. Veo3-1P aligns strongly with established image and video encoders.

with downstream performance in semantic and even non-semantic tasks (Zhu et al., 2026). In addition, we also include in Figure 4 the alignment of diffusion representations against DINOv2 (Oquab et al., 2024), a powerful image representation model, and VideoMAEv2 (Wang et al., 2023), a strong video representation model. Note that Veo3's features, especially when probed at the appropriate depth and noise levels, are significantly more aligned with text compared to Wan2.2. We emphasize that the captions are not given as context during generation. This suggests that a powerful video generative model has the capacity of encoding rich semantic information within its intermediate features. We extend our analysis to also include CogVideoX-5B Hong et al. (2022); Yang et al. (2024) and report alignment against other text encoders in appendix fig. 10. We find that the trend across noise levels is consistent across text encoders and stronger video generation models achieve higher alignment scores: Veo3-1P $>$ Wan2.2-1P $>$ CogVideoX-5B.

3.3 Linear and Attention Probes

We use linear and attention probes to further assess the quality of the diffusion representations. The linear head does global pooling over spatio-temporal dimensions followed by a linear projection. The attention probe uses a cross-attention head, with task-specific queries. We show results for video classification probes in Fig. 5.

We also experiment with lightweight decoders that reconstruct the input in RGB space given the representations from different blocks and noise levels, to get a sense of the compression and artefacts present at each stage, but the results are less conclusive; see details and visualisations in appendix D.

3.4 Discussion

Feature routing across network depth: The evolution of diffusion representations follows a dual-axis trajectory along both depth l and noise level t. Interestingly, we observe distinct evolution patterns for the two models. The results for Wan 2.2-1P (Figure 3 right, Figure 4 right) show a

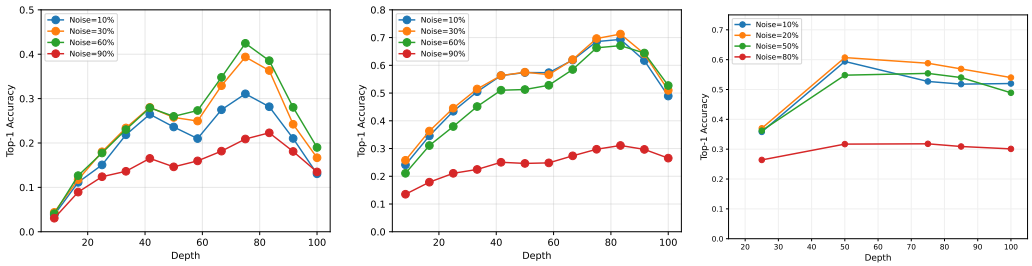


Figure 5: **Left, middle**: Linear and attention probing of Veo3-1P with SSv2 video classification. Observe that both the linear and attention probes peak at around 70-80% of the model depth and 30-60% noise level, indicating the optimal spot for extracting representations. **Right:** Attention probing of Wan2.2-1P peaks at around 50% depth and 20% noise level. Linear probing results for Wan2.2-1P are in the appendix fig. 9

unimodal pattern, aligning with broader diffusion literature, e.g., (Luo et al., 2023). The alignment peaks early, at around 25% depth, indicating that the model resolves global structure and scene details early before gradually settling finer details. Conversely, Veo3-1P exhibits a distinct bimodal pattern, a phenomenon not previously reported in the diffusion literature. This mirrors the geometry of representations learnt by large-scale transformers (Valeriani et al., 2023), where the representation manifold expands early on (first peak), contracts significantly in the middle layers to route information, and then expands again toward the end (second peak) before the final output projection. Regarding the differences between the two models, these could be explained by the differences in scale, training data, or training objectives (diffusion vs flow matching), but unfortunately we cannot provide a conclusive analysis, as ablations are prohibitively expensive to run.

60%-noise semantic bottleneck: Despite exhibiting divergent patterns across network depth, both models share a consistent optimal noise level for alignment against both text and visual encoders at $t = 60\%$. This indicates the emergence of a semantic bottleneck for both models at this noise level.

Evolution of feature complexity, linear vs. attention probes: Further analysis of Figure 5 reveals a slight shift between the optimal extraction points for linear probes ($t \approx 60\%$, $l \approx 70\%$) and attention probes ($t \approx 30\%$, $l \approx 80\%$). While both fall within the second mode of Veo3-1P's bimodal pattern, indicating the sweet spot for semantics, this shift suggests an evolution in feature complexity. At deeper blocks and lower noise levels, linear probes and zero-shot alignment metrics become insufficient possibly because semantic information becomes patch-specific and spatially scattered. Instead, attention probes excel by acting as a dynamic spatio-temporal pooling mechanism, selectively attending to and aggregating these distributed, fine-grained details. This suggests that lower-noise representations do not discard semantic meaning; they simply require more complex probes to decode it. Importantly, the superiority of the attention probe is not merely an artefact of its higher parameter count (6M vs. 1M for the linear probe). If parameter count were the sole driver, the attention head would uniformly dominate across all noise levels. Instead, at extreme noise levels ($t \approx 90\%$, red lines), its performance is severely constrained (≈15–30%), offering minimal advantage over the linear probe at early-to-mid depths. Finally, all probes experience a sharp performance drop at the deepest layers, indicating that the network discards semantic abstractions at the very end to perform pixel-level signal reconstruction.

Emergence of general-purpose features: Overall, these alignment results are remarkable. Prior evaluations of generative video models, such as WALT (Zhu et al., 2026), demonstrated limited alignment with semantic encoders. In contrast, our findings show that training purely for video generation at scale is sufficient to endow a model with powerful, general-purpose features. While these large-scale diffusion models are extensively trained with text conditioning, they are never explicitly optimised to align their internal states with text representations (unlike contrastive learning). The implicit emergence of this alignment is a significant finding not reported in prior work. Furthermore, both models demonstrate strong alignment with established visual encoders (Figure 4). Veo3-1P, in particular, exhibits high alignment with DINOv2 and VideoMAEv2, models designed for image and video understanding tasks respectively. This confirms that generative representations possess strong potential for diverse downstream tasks requiring a combination of appearance, motion, and

Model	Pre-training	Model size (M)	Top-1 accuracy (%)
VideoMAEv2-g (Wang et al., 2023)	MAE	1,013	65.6
VideoPrism-g (Zhao et al., 2024)	MAE, Contrastive	1,113	65.4
4DS-j (Carreira et al., 2025)	MAE	21,495	68.2
InternVideo2 (Wang et al., 2024b)	MAE, Contrastive, Captioning	6,000	67.7
V-JEPA-H (Bardes et al., 2024)	Masked feature prediction	635	72.2
V-Walt (Vélez et al., 2025)	Diffusion	1,900	59.7
Wan2.2-1P (Our)	Diffusion	∼7,480	60.8
Veo3-1P (Our)	Diffusion	-	71.3
Veo3-1P w/ data augm. (Our)	Diffusion	-	**72.6**

Table 1: Comparison with state-of-the-art on SSv2 video classification. These methods share the same evaluation protocol: frozen representations are fed into a trained one-block attention read-out.

semantic understanding. Finally, the consistently higher alignment scores of Veo3-1P compared to Wan 2.2-1P confirm that stronger generative models yield correspondingly richer and more powerful representations.

Combination of features: We investigated whether combining features across multiple stages of the diffusion process could yield complementary information. We include details in Appendix B.

4 Diffusion models as video encoders

We demonstrate the potential of diffusion models to act as general-purpose video encoders by evaluating frozen diffusion representations on diverse semantic and non-semantic downstream tasks. While our latent space analysis in Section 3 confirmed the presence of optimal extraction points across both open-weight (Wan 2.2) and proprietary (Veo3) models, it also suggests that feature alignment scales directly with generative capability. Therefore, to test our hypothesis and set a strong video diffusion based baseline, we focus our downstream evaluations on the highly scaled Veo3 model. We rely on one-block attention decoders to map diffusion representations to various task embedding spaces, similar to 4DS (Carreira et al., 2025). For captioning, we use a language model fine-tuned on top of the frozen visual diffusion tokens. Only these lightweight decoders are trained, the generative model stays frozen, preserving its generation capabilities.

4.1 Semantic understanding

Video classification: We use the SSv2 dataset (Goyal et al., 2017) for this task and follow the protocol in (Carreira et al., 2025), using simple linear and one-block attention decoders on top of frozen diffusion representations. The SSv2 dataset contains 220,847 short videos (2-6s long), sampled at 12fps, representing 174 classes. The videos depict actions that differ in fine motion-related details, requiring temporal understanding, e.g. *pouring something into something* vs *pretending to pour something into something*.

To reduce computation cost, for Veo3-1P experiments, we do not do data augmentation on the raw videos like in (Carreira et al., 2025). Instead, we process the original videos and we apply data augmentations directly on the intermediate activations: temporal masking (i.e. set to 0) of 16.7% of the frames, 40% dropout in the attention mask and label smoothing of 0.2. For Wan2.2-1P experiments, we apply data augmentation as in (Carreira et al., 2025). We report performance in table 1, compared to strong competitor models. Veo3-1P achieves SOTA performance on this task in this evaluation setup. Note that even without the data augmentation, Veo3-1P outperforms all baselines, except V-JEPA.

Image and video captioning: We use the Gemma2 (2B) (Gemma Team, 2024) text-only language model as a decoder for image and video captioning. We connect diffusion representations to Gemma2 in two steps: (1.) project the diffusion representation into 32 learnable query tokens (Q) using a cross-attention adapter, inspired by the decoder design of (Sajjadi et al., 2022), with three cross-attention blocks; and (2.) prepend Q before the [BOS] token and process the entire sequence with the LLM, using teacher forcing during training and next-token prediction during inference. We train separate decoders for each of the MS-COCO

Model	SSv2		COCO (Test)		VATEX (Test)		VATEX [Frozen LLM] (Test)	
	CIDEr	BLEU@4	CIDEr	BLEU@4	CIDEr	BLEU@4	CIDEr	BLEU@4
SigLIP-so400m/14	204.5 ± 18.7	34.5 ± 1.8	$\mathbf{118.5 \pm 0.5}$	$\mathbf{33.2 \pm 0.2}$	$\mathbf{66.0 \pm 0.8}$	$\mathbf{34.0 \pm 0.3}$	$\mathbf{64.5 \pm 0.8}$	$\mathbf{32.9 \pm 0.2}$
SigLIP2-B/16	198.6 ± 5.8	33.9 ± 0.6	114.2 ± 2.3	32.4 ± 0.5	58.4 ± 0.8	31.4 ± 0.4	56.2 ± 0.9	30.5 ± 0.1
Veo3-1P @ 30%	$\mathbf{289.5 \pm 7.7}$	$\mathbf{45.5 \pm 0.7}$	54.9 ± 0.8	20.2 ± 0.2	44.8 ± 1.5	28.1 ± 0.5	40.2 ± 0.4	26.5 ± 0.2
+ Noise Aug.	280.4 ± 16.5	44.8 ± 1.7	69.3 ± 1.0	23.5 ± 0.3	56.7 ± 0.1	32.3 ± 0.1	48.5 ± 0.5	29.6 ± 0.2
+ High res.	-	-	102.0 ± 1.3	30.4 ± 0.4	-	-	-	-

Table 2: Captioning results across SSv2, COCO, and Vatex Datasets reporting the CIDEr (↑) and BLEU@4 (↑) scores' mean and standard error across 5 seeds. Higher is better.

image captioning dataset (COCO) (Chen et al., 2015), the SSv2 action classification dataset[1] and the VATEX dataset (Wang et al., 2019). Table 2 shows CIDEr and BLEU@4 metrics compared against two baseline models: SigLIP-so400m/14 model from the PaliGemma series (Beyer et al., 2024; Steiner et al., 2024) and the SigLIP2-B/16 model (Tschannen et al., 2025).

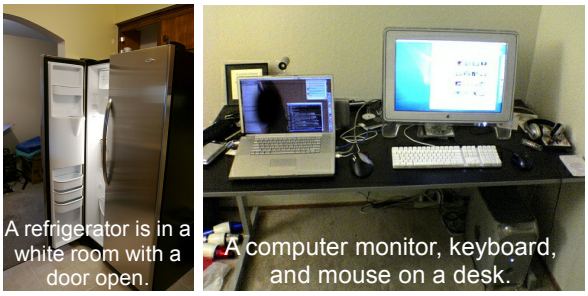


Figure 6: Veo3-1P generated captions.

Our setup does **not** involve tuning the adapter and LLM decoder on large vision-language data, so it is not directly comparable to SoTA methods like BLIP-2 ((Li et al., 2023b)) and PaLI Gemma ((Beyer et al., 2024)). We train our adapter and LLM only on the train splits of the considered datasets to probe the existence of semantic knowledge decodable through an LLM. We observe that video diffusion representations excel on SSv2 but struggle on COCO and VATEX, compared to the SigLIP baselines, which are highly optimised for alignment with text. Nonetheless, the model produces reasonable captions, as shown on the COCO dataset in Fig. 6.

VATEX [Frozen LLM]: We freeze the LLM in this experiment and train only the cross-attention adapter on the VATEX train split. Results are reported in the last two columns of Table 2. Model performance decreases only slightly relative to full fine-tuning. Moreover, we significantly outperform VideoPrism which reports a CIDEr of 31.7 on this dataset. Although VideoPrism does not fine-tune their adapter on captioning, they pre-train it on a large vision-language corpus.

Data augmentation using multiple noise levels: As we do not pre-train our adapter / LLM on a large corpus of image-text pairs, we address over-fitting as follows. We augment the data by concatenating the datasets across noise levels $10\%, 30\%, 60\%$ and 90%. We find that for COCO and VATEX using the 10%, 30% and 60% noise injection levels is optimal. The model is evaluated on the 30% level in all cases. See Table 2 under *'+ Noise Aug.'*. For SSv2, we do not see any improvements with data augmentation, likely because the caption space is simpler and the training dataset is much larger.

Benefits of high-resolution: COCO images contain several small objects and the model needs to mention a few of them to score highly on the CIDEr metric. We ablate doubling the input resolution. We pick the optimal combination of noise levels ('+ Noise Aug.') for each resolution on the validation split and report test set numbers in Table 2 under *'+ High res.'*. The high res. model achieves gains $+32.7$ CIDEr points. Thus by using multiple noise levels as data augmentation and using higher resolution diffusion representations, one can reduce the gap to discriminative models trained explicitly for the captioning task on large-scale vision-language datasets.

4.2 Geometry Understanding

Depth estimation: We evaluate monocular depth prediction on ScanNet (Dai et al., 2017). Following prior work (Carreira et al., 2025), we report the absolute relative error (AbsRel), computed as $|d^* - d|/(d + \epsilon)$ where d^* and d denote predicted and ground-truth depth, respectively, and the

[1]The SSv2 dataset consists of templated class labels like 'putting something into something'. We use the labels that incorporate the nouns as well, e.g. 'putting an apple into a fruit basket'. We prepend this with 'A person is' to form a full sentence: 'A person is putting an apple into a fruit basket'.

threshold accuracy δ_1. Target depth values outside the $(0.001, 10)$ m range are masked out. We use a Dense Prediction Transformer (DPT) readout head (Ranftl et al., 2021) with approximately 23M trainable parameters, trained with the scale-invariant logarithmic (SiLog) loss (Eigen et al., 2014). The entire diffusion backbone remains frozen.

Model	Size (M)	AbsRel (↓)
Veo3-1P (Our)	-	**0.075**
VideoPrism-g Zhao et al. (2024)	1,113	0.096
V-JEPA-H Assran et al. (2025)	635	0.102
VideoMAE-H Tong et al. (2022)	633	0.103
VideoMAEv2-g Wang et al. (2023)	1,013	0.108
VideoMAE-L Tong et al. (2022)	305	0.109
V-JEPA-L Assran et al. (2025)	307	0.114
VideoPrism-B Zhao et al. (2024)	114	0.120
Wan2.2-1P (Our)	$\sim$ 7,480	0.134
SigLIP-2B Chen et al. (2023)	1,705	0.147
VideoMAE-B Tong et al. (2022)	87	0.151

Table 3: ScanNet depth estimation (AbsRel).

Model	Category	Peak R^2
DINOv2-B Oquab et al. (2024) (L7)	SSL	**0.677**
DUSt3R (L18) Wang et al. (2024a)	Geometry	0.596
Veo3-1P (Our)	Diffusion	0.589
MASt3R (L19) Leroy et al. (2024)	Geometry	0.575
DINOv3 (L7) Siméoni et al. (2025)	SSL	0.561
Perception Encoder (L15) Bolya et al. (2025)	Contrastive	0.534
VGGT-DINOv2 (L16) Wang et al. (2025)	Geometry	0.523
V-JEPA 2.1 (L5) Mur-Labadia et al. (2026)	SSL	0.523
SD 2.1 (L1) Rombach et al. (2022)	Diffusion	0.537
CroCo (L11) Weinzaepfel et al. (2022)	Geometry	0.451
SVD Blattmann et al. (2023)	Diffusion	0.409
CLIP-B (L8) Radford et al. (2021)	SSL	0.402

Table 4: Camera motion decodability

Figure 13 (left) presents AbsRel across all blocks and noise levels in Veo3-1P. The results confirm the sweet spot for extracting representations around (depth 80%, noise 30%), identified using zero-shot alignment and classification probes. The best configuration achieves AbsRel of **0.075** and δ_1=0.952, a 10.7% relative improvement over the frozen-feature baseline of 0.084 reported in (Carreira et al., 2025). To our knowledge, this is the best result on ScanNet depth with frozen video model features, see table 3 for a comparison where we report the best performing block and noise level for Veo3-1P and Wan2.2-1P. Figure 13 (middle) maps depth estimation performance for Wan2.2-1P which peaks at noise 20% but a much lower depth of 50%, matching the optimal setting for SSv2 action classification.

Camera motion decodability: To evaluate whether diffusion natively captures intrinsic 3D geometry, we benchmark on the Poincaré Task (M4 metric) introduced in SeeSE3 (Chen et al., 2026). Inspired by the philosophical premise that a passive observer cannot comprehend space without active movement, this task tests whether a model naturally organises its latent manifold to reflect 3D Euclidean transformations ($SE(3)$). Specifically, it uses a lightweight Siamese MLP to measure whether physical camera motion can be linearly decoded directly from the displacement between two latent features. Remarkably, representations extracted from Veo3-1P exhibit profound spatial awareness; see results in Tab. 4. Veo3-1P significantly outperforms dedicated self-supervised video architectures like V-JEPA, CroCo, even MASt3R (trained explicitly for geometry), and achieves decodability scores that closely rival DINOv2 and DUSt3R. These results prove that the generative objective inherently builds geometric representations without explicit 3D supervision. We also train an attention probe for standard camera pose estimation, which reveals competitive performance with DINOv2; see details in appendix E and Figure 13 right.

5 Conclusion

We show that frontier video diffusion models can act as competitive video encoders on a diverse set of tasks (video classification, depth estimation, camera pose estimation, video captioning), and that generation power correlates with understanding capabilities. Using zero-shot probes and trained lightweight decoders, we investigated the structure of the diffusion latent space and identified the optimal block and noise level to extract representations for understanding tasks. Once the optimal (block, noise level) identified, we only need to run a single forward pass through the model, being as efficient as large-scale discriminative visual encoders. As future work, we aim to study how to compress the diffusion features or identify optimal subspaces for various tasks to improve performance, e.g. on captioning tasks.

Limitations: Our SOTA results are obtained with a proprietary video model, Veo3, with analysis extended to open-source models (Wan 2.2, CogVideoX-5B), limiting the reproducibility of our study. However, we believe that this is a timely and useful investigation for the community and we hope it will encourage more work in studying large-scale diffusion models for representation learning.

AI USE STATEMENT

In this work, we used generative AI tools for: implement methods, clean and reformat dataset, support qualitative and thematic data analysis.

We have not used generative AI tools for: Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, interpret results.

Additionally, we used generative AI tools for: create or modify scientific figures or images, create or edit software code, summarize or analyse existing literature, sourcing/searching for information, identify relevant literature

We have reviewed all AI-assisted work. We manually reviewed all generated code which was used to run experiments or create plots for the paper. We manually checked existing literature to ensure our understanding of prior works is correct.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Several experiments in this manuscript study the closed-weights Veo3 model. To assist with the reproducibility of our findings, we have analysed open-weights Wan2.2 and CogVideoX-5B as well.

Our experiments evaluate the activations of these video diffusion backbones on downstream tasks. The datasets for these downstream tasks are publicly available (SSv2, COCO, VATEX, ScanNet). The text encoders used for alignment are open-weight models (Gemma2-9B-it, LLaMA-65B, OpenLLaMA-13B). The text decoder used for captioning is the open-weight Gemma2-2B.

We detail any hyper-parameters not already discussed in the main paper in the appendix in sections F and G.

ACKNOWLEDGMENTS

To be written in the camera ready.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *TMLR*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2. Featured Certification.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, et al. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

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.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network, 2025. URL https://arxiv.org/abs/2504.13181.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge (Jimmy) Shi, et al. Genie: generative interactive environments. In *ICML*. JMLR.org, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

João Carreira, Dilara Gokay, Michael King, Chuhan Zhang, Ignacio Rocco, Aravindh Mahendran, Thomas Albert Keck, et al. Scaling 4d representations, 2025. URL https://arxiv.org/abs/2412.15212.

Caroline Chen, Sayna Ebrahimi, Fedor Kitashov, Ming-Hsuan Yang, Leonidas Guibas, Viorica Pătrăucean, and Maks Ovsjanikov. SeeSE3: The emergence of 3d space in vision features. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=iW8SfqJzsb.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. PaLI-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

Sander Dieleman. Diffusion is spectral autoregression. sander.ai, September 2024. URL https://sander.ai/2024/09/02/spectral-autoregression.html. Accessed: 2026-04-17.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, pp. 2366–2374, 2014.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Google DeepMind. Veo 3 technical report. Technical report, Google, 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

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.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *NeurIPS*, 2020.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *ECCV*, 2023. URL https://api.semanticscholar.org/CorpusID:266163109.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, pp. 2451–2463, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8037):647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 15979–15988, 2021. URL https://api.semanticscholar.org/CorpusID:243985980.

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

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *ICML*, volume 235, pp. 20617–20642. PMLR, 2024.

Hyeonho Jeong, Chun-Hao P. Huang, Jong Chul Ye, Niloy J. Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *CVPR*, pp. 7276–7287, June 2025.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pp. 3128–3137, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R, 2024.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, pp. 2206–2217, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *NeurIPS*, 2023. URL https://openreview.net/forum?id=Vm1zeYqwdc.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning, 2026. URL https://arxiv.org/abs/2603.14482.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010. URL https://www.cs.toronto.edu/~fritz/absps/reluICML.pdf.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joseph Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14386–14397, 2023. URL https://api.semanticscholar.org/CorpusID:266174654.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, et al. Genie 2: A large-scale foundation world model. https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/, 2024. Accessed: 2026-09-26.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pp. 12179–12188, 2021.

Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altche, Michal Valko, Jean-Bastien Grill, Aaron van den Oord, and Andrew Zisserman. Broaden Your Views for Self-Supervised Video Learning. In *ICCV*, pp. 1235–1245, 2021. doi: 10.1109/ICCV48922.2021.00129. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00129.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, June 2022.

Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *CVPR*, pp. 6229–6238, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. PaliGemma 2: A Family of Versatile VLMs for Transfer. *arXiv preprint arXiv:2412.03555*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *NeurIPS*, 2023.

Pedro Vélez, Luisa F. Polanía, Yi Yang, Chuhan Zhang, Rishabh Kabra, Anurag Arnab, and Mehdi S. M. Sajjadi. From image to video: An empirical study of diffusion representations. In *ICCV*, pp. 16948–16958, October 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, pp. 5294–5306. IEEE, 2025.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE v2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *CVPR*, 2024a.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pp. 4581–4591, 2019.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, pp. 396–416. Springer-Verlag, 2024b. ISBN 978-3-031-73012-2. doi: 10.1007/978-3-031-73013-9_23. URL https://doi.org/10.1007/978-3-031-73013-9_23.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised pre-training for 3d vision tasks by cross-view completion. *NeurIPS*, 35:3502–3516, 2022.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one D4RT at a time. In *CVPR*, 2026.

Long Zhao, Nitesh Bharadwaj Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming-Hsuan Yang, David A Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. VideoPrism: A foundational visual encoder for video understanding. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60785–60811. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhao24f.html.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025. URL https://openreview.net/forum?id=SI2hI0frk6.

Tyler Zhu, Tengda Han, Viorica Pătrăucean, Leonidas Guibas, and Maks Ovsjanikov. Dynamic reflections: Probing video representations with text alignment. In *ICLR*, 2026. URL https://openreview.net/forum?id=gE17TwVMNh.

A MUTUAL K-NN ALIGNMENT METRIC

We describe the Mutual k-Nearest Neighbours (MkNN) alignment metric used throughout the paper, following (Huh et al., 2024; Zhu et al., 2026).

Notation. Consider a dataset of N videos $\{v_i\}_{i=1}^N$, each paired with some text c_i. We compare the representations of two encoders whose embedding dimensionalities may differ.

Diffusion encoder. To use the LDM f_θ as a video encoder, we run the model in its forward (training) mode at a chosen noise level t and extract the intermediate activations $h_t^{(l)} \in \mathbb{R}^{F' \times H' \times W' \times D}$ from layer l of the backbone, where F', H', W' are the spatiotemporal dimensions of the latent and D is the channel dimension. We aggregate these activations via global average pooling over the spatiotemporal axes to obtain a single embedding vector per video:

$$\mathbf{x}_i = \mathrm{AvgPool}\big(h_t^{(l)}(v_i)\big) \in \mathbb{R}^D. \tag{1}$$

We stack all video embeddings into a matrix $X \in \mathbb{R}^{N \times D}$.

Reference encoder. A reference encoder E_{ref} (e.g. a text encoder applied to text descriptions, or an independently trained video encoder) maps each input to an embedding of potentially different dimensionality D':

$$\mathbf{y}_i = E_{\mathrm{ref}}(c_i) \in \mathbb{R}^{D'}, \quad Y \in \mathbb{R}^{N \times D'}. \tag{2}$$

When the reference encoder is a pure image model (e.g. DINOv2), we average its frame-level features across the temporal dimension, following (Zhu et al., 2026).

Feature preprocessing. Before computing nearest neighbours, we clip each feature matrix to suppress outliers: for a chosen quantile q (we use q=0.95), we compute $\tau = \mathrm{Quantile}_q\big(|X|\big)$ and clip all values to $[-\tau, \tau]$. We then ℓ_2-normalise each row so that similarities reduce to dot products.

Nearest-neighbour graphs. Let $\mathcal{N}_k^X(i)$ denote the set of k nearest neighbours of sample i in the embedding space X, computed via dot-product similarity (excluding self-matches). We define an analogous set $\mathcal{N}_k^Y(i)$ for space Y. Equivalently, we can encode these as binary indicator matrices $M^X, M^Y \in \{0,1\}^{N \times N}$, where

$$M_{ij}^X = \begin{cases} 1 & \text{if } j \in \mathcal{N}_k^X(i), \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Alignment score. The MkNN alignment between X and Y is the mean fraction of shared neighbours:

$$\mathcal{A}_{\mathrm{M}k\mathrm{NN}}(X, Y) = \frac{1}{kN} \sum_{i=1}^N \sum_{j=1}^N \big(M^X \odot M^Y\big)_{ij} = \frac{1}{N} \sum_{i=1}^N \frac{\big|\mathcal{N}_k^X(i) \cap \mathcal{N}_k^Y(i)\big|}{k}, \tag{4}$$

where $\odot$ denotes element-wise multiplication. This score ranges from 0 (no overlap) to 1 (perfect agreement) and is invariant to the dimensionality and scale of each space, since it depends only on the rank-order of pairwise similarities.

Layer optimisation. Both the diffusion backbone and multi-layer reference encoders expose representations from multiple intermediate layers. Following (Huh et al., 2024), we sweep over all pairs of layers (l, l')—where l indexes a layer in f_θ and l' a layer in E_{ref}—and report the pair that maximises $\mathcal{A}_{\mathrm{M}k\mathrm{NN}}$:

$$(l^\star, l'^\star) = \arg\max_{l, l'} \ \mathcal{A}_{\mathrm{M}k\mathrm{NN}}\big(X^{(l)}, Y^{(l')}\big). \tag{5}$$

We use k=10 in all experiments.

B COMBINING VEO3 LATENTS ACROSS BLOCKS AND NOISE LEVELS

In this section, we provide details and preliminary results regarding the experiments mentioned in Section 3.4 on combining Veo3 representations across different blocks and noise levels.

B.1 EXPERIMENTAL SETUP AND ARCHITECTURES

To investigate the complementarity of Veo3 representations, we extract features from 12 equally spaced blocks across 4 noise levels ($t \in \{10\%, 30\%, 60\%, 90\%\}$). Features are averaged across space and time. This results in $4 \times 12 = 48$ distinct D-dimensional feature vectors per video. We evaluate four architectures to map these 48 feature vectors into a single global video representation:

- **Linear adapter**: This computes a simple weighted sum of the 48 feature vectors:

$$F_{combined} = \sum_{i=1}^{48} w_i \cdot F_i$$

  where $w \in \mathbb{R}^{48}$ is a vector of learnable scalar weights. This architecture has the fewest parameters (48) and demonstrated the best generalization.
- **Shared MLP**: The input features are transposed to $(D, 48)$ and passed through a series of linear layers with ReLU Nair & Hinton (2010) activations shared across the feature dimension:

$$\begin{aligned}\text{Input}(48, D) &\to \text{Transpose}(D, 48) \to \text{Linear}(48 \to M) \to \text{ReLU} \\ &\to \text{Linear}(M \to H) \to \text{ReLU} \to \text{Linear}(H \to 1)\end{aligned}$$

  The output is then squeezed to produce the final $1 \times D$ vector. We experimented with configurations such as $M = 8, H = 8$.
- **Self-attention**: A learnable CLS token $c \in \mathbb{R}^D$ is concatenated with the 48 features to form a $49 \times D$ matrix. All 49 tokens attend to each other via standard self-attention, with queries, keys, and values projected to $d_k = 64$. The output corresponding to the CLS token is extracted and projected back to D.
- **Cross-attention**: A Perceiver-style architecture where a single learnable query token cross-attends to the 48 features (which are first compressed to $d_k = 64$). This significantly reduces parameters compared to full self-attention while allowing dynamic weighting.

B.2 TRAINING OBJECTIVES

We considered two primary training losses to train these adapters:

- **Cross-entropy loss** (for direct classification on SSv2):

$$\mathcal{L}_{CE} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{c=1}^{C} \mathbf{1}[y_i = c] \log \frac{\exp(\hat{y}_{i,c})}{\sum_{c'=1}^{C} \exp(\hat{y}_{i,c'})}$$

  where B is the batch size, C is the number of classes (174 for SSv2), y_i is the true label, and $\hat{y}_{i,c}$ are the logits produced by adding a linear classification head after the adapter.
- **Multi-positive InfoNCE loss** (for cross-modal alignment):

$$\mathcal{L}_{InfoNCE} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\sum_{j \in \mathcal{P}(i)} \exp(\text{sim}(\mathbf{v}_i, \mathbf{v}_j)/\tau)}{\sum_{k=1}^{B} \exp(\text{sim}(\mathbf{v}_i, \mathbf{v}_k)/\tau)}$$

  Here, $\mathbf{v}_i$ is the L_2-normalized adapter output for sample i, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature parameter. The set $\mathcal{P}(i)$ contains the indices of the k-nearest neighbours in the text feature space (e.g., Gemma embeddings), acting as multiple positives.

B.3 RESULTS

Dataset. We run preliminary experiments using a subset of the SSv2 training set. Specifically, we keep 100 training samples per class, resulting in a training set with 17400 samples. We evaluate on the full SSv2 validation set. The baseline performance of a linear classifier trained on the single best block and noise level using this subset is 17.04%. This is a modest performance, but it is significantly better than chance, providing the necessary signal for our investigation; see Table 5.

Adapter	Training data	Training objective	Accuracy
Best single block (baseline)	–	–	17.04
Linear	VATEX	Text alignment	21.4
Linear	SSv2	Text alignment	21.9
Cross-attention	SSv2	Classification	**24.9**

Table 5: SSv2 classification (linear readout head, no augmentation) with Veo3 features. "Best Single block" uses the single optimal block and noise level combination under this classification protocol. "Combined" uses a learned adapter to fuse features across 12 blocks and 4 noise levels. Note that training for alignment with text captions (i.e., *without any* SSv2 class supervision) still leads to improvement in downstream classification with a linear adapter. Training a cross-attention based adapter directly for SSv2 classification yields the best result overall.

Linear adapters consistently outperform non-linear ones in terms of zero-shot generalization across datasets and tasks. For instance, a linear adapter pre-trained purely for text alignment on VATEX generalised effectively to SSv2 out-of-the-box. Conversely, non-linear models—such as the cross-attention and self-attention architectures—were significantly harder to train. In early experiments, they frequently achieved high alignment scores on the training set while generalizing poorly to downstream classification tasks, indicating severe over-fitting.

Training for non-linear adapters. To stabilize the training of higher-capacity non-linear adapters, we found that modifying the batch composition was critical. Specifically, increasing the effective batch size and ensuring a robust ratio of anchors, positives, and negatives prevented the cross-attention models from collapsing. Our optimal configuration utilized 16 anchors per batch. For each anchor, we sampled 10 positive matches and 22 negative matches (yielding 32 samples per anchor). This structured batching was the key heuristic that allowed the cross-attention adapters to converge without over-fitting to the training distribution.

Peak performance via cross-attention. While linear adapters are robust and easily generalisable, the carefully tuned cross-attention adapter ultimately yielded the highest peak performance when trained directly on the target distribution. When trained specifically for classification on the SSv2 dataset using the aforementioned batching heuristics, the cross-attention adapter achieved a downstream classification accuracy of 24.9% on this SSv2 subset. This represents a substantial improvement over the 17.04% baseline (which utilises a single optimal block and noise level) and also outperforms the best linear adapter's peak of 21.9% on this SSv2 subset. All of these results were obtained using a simple *linear* readout head without any data augmentation.

Correlation across training objectives. Throughout our ablation studies across different training objectives (classification, class-vector alignment, and text-embedding alignment), we observed a highly consistent empirical trend: training the adapter for classification utilising a discrete label signal simultaneously improved all proxy metrics for feature quality.

Most notably, adapters trained using a standard cross-entropy classification loss consistently exhibited *better* zero-shot cross-modal text alignment (e.g., higher mutual k-NN alignment with Gemma text features) than adapters that were explicitly trained to maximize text alignment. Similarly, 1-NN retrieval accuracy also peaked under the classification objective. We hypothesize that the discrete cross-entropy objective provides a cleaner, more stable gradient signal for the adapter to learn the underlying semantic structure of the VEO representations, compared to the noisy gradients inherently present in multi-positive contrastive alignment over text embeddings.

C Additional results with Wan2.2 and CogVideoX-5B

Due to space constraints we report additional results here in the appendix.

Shape-class purity: Figure 7 shows an example image from the shape-texture dataset (Geirhos et al., 2019), where style transfer techniques are used to inpaint objects of various classes with



Figure 7: Example image from the shape-texture bias dataset (Geirhos et al., 2019). From left to right: Original shape image, original texture image, shape and texture combined.

textures belonging to different classes, e.g. a cat shape inpainted with an elephant texture. The dataset contains 1,280 images across 16 classes (airplane, bear, bicycle, bird, boat, bottle, car, cat, chair, clock, dog, elephant, keyboard, knife, oven, truck), with 80 images per class, 10 base shape exemplars per object class (160 shapes total across the dataset). Each shape exemplar is paired with a subset of textures. There are 15 adversarial texture classes.

Figure 8 left shows Wan2.2-1P shape-purity k-nn clustering results: for every shape image, we retrieve its k nearest neighbours in the embedding space and count the number of nearest neighbours whose shape class matches the class of the reference image. Compared to Veo3-1P results (fig. 3 left), we see a lower shape bias in Wan2.2-1P. This is consistent with the alignment and classification results, which consistently show what Veo3 has stronger representations compared to Wan2.2, correlating with their generation capabilities.

SSv2 classification: Figure 9 shows linear classification for Wan2.2-1P . The peak performance at noise 50% differs slightly from the peak at 20% noise when using attention probes in figure 5 (right).

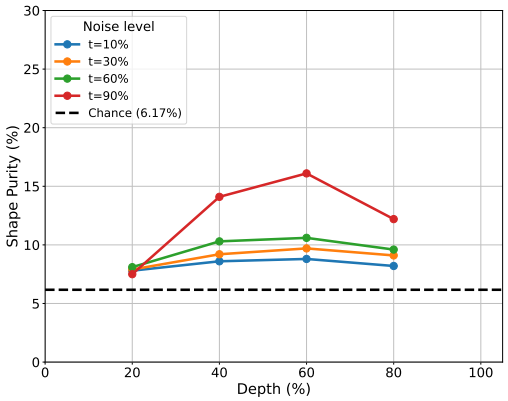


Figure 8: Shape-class purity for Wan2.2-1P.

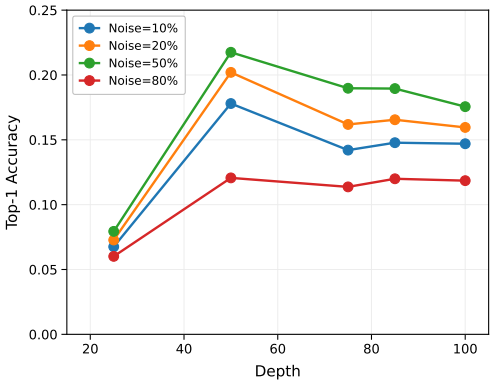


Figure 9: Wan2.2-1P linearly probed on the SSv2 action classification dataset. Performance peaks at noise 50% and depth 50%.

Alignment with text: In figure 10, we report mutual k-NN based alignment against text encodings from two other text encoders, LLaMA-65B and OpenLLaMA-13B. Additionally we evaluate CogVideoX-5B representations as well. Each data point plots the value corresponding to the optimal network depth. The pattern is consistent across text encoders, demonstrating that our conclusions are not an artifact of the specific text encoder chosen. CogVideoX-5B's alignment peaks at low noise levels and has a different pattern than its stronger video generation counterparts.

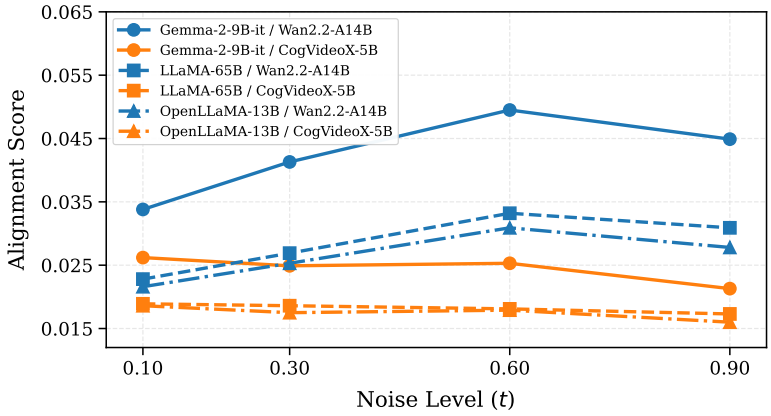
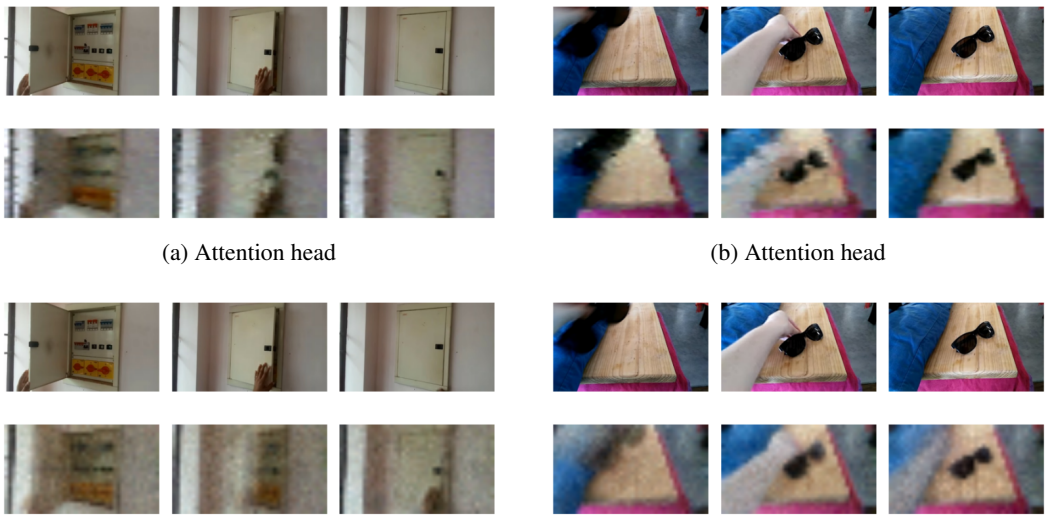


Figure 10: Extended alignment results including other text encoders and the CogVideoX-5B video diffusion model. We report the peak performance across network depths at each noise level.



(a) Attention head

(b) Attention head

(c) Linear head

(d) Linear head

Figure 11: Qualitative video reconstruction results on two example videos from the Something-Something V2 validation set. In each sub-figure, we show the ground-truth on the top row, and the reconstructions from the linear and single-layer attention head, respectively, on the bottom row. Even with these very simple readout heads, the reconstructions convey well the main elements in the scene.

Depth estimation. Fig. 13 (left and middle) presents results on ScanNet depth estimation across network depths and noise levels. We compare against the SoTA 4DS model shown as dotted-line reference. Veo3-1P outperforms the reference baseline, achieving state-of-the-art on ScanNet depth estimation in this setting. Wan2.2-1P shows poorer performance in comparison, but has the same trend as with SSv2 attention probing peaking at 20% noise and 50% depth.

D PREVIEW DECODING

We train linear and attention preview decoders on the Something-Something-V2 dataset (Goyal et al., 2017), i.e. we attempt to decode early (and efficiently) the final RGB output to get a qualitative glimpse into the features encoded at different blocks and noise levels. We use the mean L2 error on RGB values as the training loss.

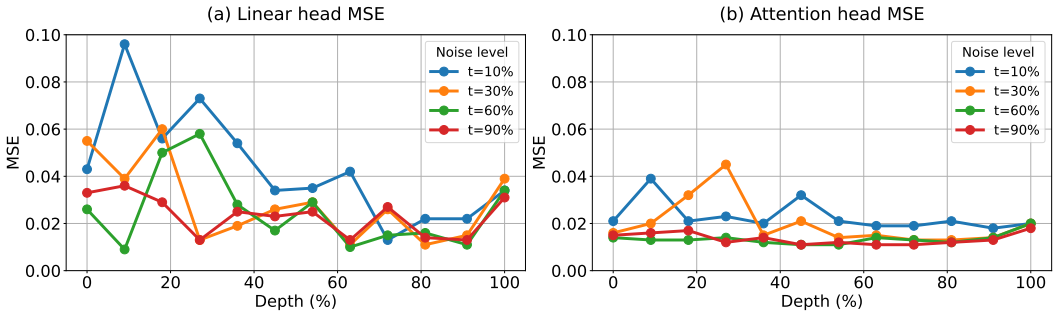


Figure 12: Evaluating RGB readout heads attached at varying network depths and noise levels. Left: linear probe. Right: attention probe.

Linear readout head: The linear head applies a single learned affine projection from the token embedding space to RGB colour. Given input tokens $\mathbf{X} \in \mathbb{R}^{B\times F\times N\times C}$ (B is the batch size, F is the number of frames, N is the number of spatial tokens per frame, and C is the number of channels per token), the head computes $\hat{\mathbf{Y}}_{\text{lin}} = \mathbf{X}W + b$ with $W \in \mathbb{R}^{C\times 3}$ and $b \in \mathbb{R}^3$. The output is reshaped to $(B, F, H, W, 3)$, where $H \cdot W = N$, recovering the native spatial layout of the Veo3 token grid.

Attention readout head: The attention head uses cross-attention with learnable queries to decode a target-resolution RGB video. A bank of $F \cdot H \cdot W$ query vectors is learned, each of dimension $d < C$. The input tokens are first flattened to $(B, F\cdot N, C)$ and linearly projected to keys and values of dimension $d = 512$. A standard multi-head dot-product attention layer with h=8 heads (head dimension 64) then computes cross-attention from the learned queries to the projected keys and values, yielding an output of dimension d=512 per query. A final linear projection maps each query output to 3 (RGB) channels, and the result is reshaped to $(B, F, H, W, 3)$.

Both heads are trained with per-pixel L2 loss and evaluated with MSE. The optimizer is Adam with a linear warm-up $(1,000$ steps) to a peak learning rate of 10^{-4}, with gradient accumulation to an effective batch size of 16.

Results. Figure 11 presents frames from the original video and their previews decoded from the outputs at depth 75%, evaluated at noise level 60%. The linear decoder presents frames that are sharper, but temporally misaligned with the ground truth, as compared to the attention head.

Figure 12 shows MSE reconstruction metrics for attaching the preview head at various network depths, and training it on various levels of injected noise. Unlike the zero-shot alignment scores or the classification probes, regressing the denoised video's pixel values from various heads does not show a clear pattern with an optimal sweet spot for extracting features. The linear head does show generally lower error at higher network depths, reflecting that deeper layers are closer to the final latents to be decoded into RGB values. On average, the attention head outperforms the linear head by having better temporal alignment with the original video.

E CAMERA POSE ESTIMATION

We probe the ability to infer the 6DoF relative camera poses from diffusion representations. We use the setup and metric introduced in (Carreira et al., 2025), applied to ScanNet dataset. Given diffusion representations for a clip of F frames, we train a one-block attention decoder to predict the relative pose between the first and last frame of the clip, in the form of a $12D$ vector encoding an SE(3) pose transformation (3×3 estimated rotation matrix and a 3×1 translation vector). As evaluation metric, we use end-point-error (EPE), which considers the rotation and translation components jointly; see (Carreira et al., 2025) for more details. We obtain 1.10 EPE, which is on par with a strong DINOv2 baseline obtaining 1.08 EPE. The sweep across blocks and noise levels (Figure 13 left) is consistent with the findings for the other tasks, with performance peaking reliably in the mid-to-late stages of the network ($\sim$75% depth), at medium noise level (60%).

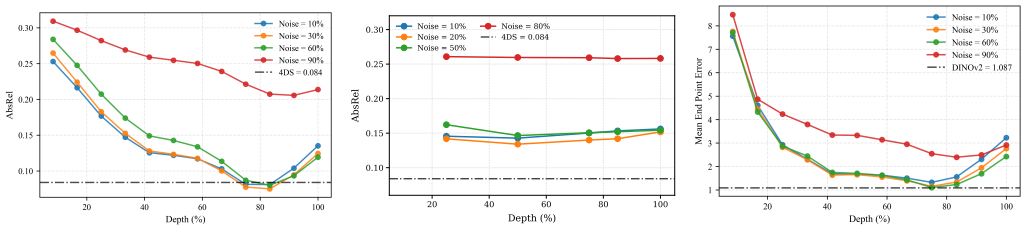


Figure 13: Performance on geometry tasks tasks across blocks and noise levels. **Left:** Veo3-1P probed for depth estimation on ScanNet (AbsRel ↓). **Middle:** Wan2.2-1P probed for ScanNet depth estimation. **Right:** Veo3-1P probed for camera pose estimation EPE on ScanNet. The dashed lines mark strong baselines: 4DS (Carreira et al., 2025) and DINOv2 on depth estimation and camera pose estimation respectively.

F Optimizer hyper-parameters

Action classification We train on the Something-something-V2 dataset for $40,000$ steps using the AdamW optimizer Loshchilov & Hutter (2017) with a weight decay of 10^{-4}. We use a cosine learning rate decay with a linear warm-up to 3×10^{-4} over $1,000$ steps. It decays down to 10^{-7}. Batch size is 32.

For Wan2.2-1P experiments, 16 frames are indexed with stride 2 and resized to 256×256. For Veo3-1P experiments, the activations that we obtained were such that the last frame is repeated to fit the Veo3 input shape of 192 frames.

ScanNet depth estimation We again train with a batch size of 32 for 40,000 steps. The peak learning rate is 3×10^{-5}, warmed-up over $1,000$ steps and then decaying down with a cosine decay rate to 10^{-7}. The AdamW weight decay is set to 10^{-6}. We random crop a window of size 224×224 and apply random flip augmentation during training.

G Image and video captioning details

The cross-attention adapter alternates cross-attention layers from learned queries into input vision tokens with self-attention layers across query tokens. The adapter is designed to output $\mathbf{R}^{2304}$ features matching Gemma2-2B's input embedding size. Gemma2 is provided an attention mask that is bi-directional on the vision tokens and causal on the text tokens. The loss is applied to the text tokens **after** the [BOS] token. During inference we use greedy sampling and discard all tokens after the first [EOS] token is sampled. A maximum sampling sequence length of 64 tokens is used for the COCO and SSv2 datasets and 96 tokens are sampled for the VATEX dataset. This was adjusted for VATEX to accommodate longer captions in the dataset.

For SigLIP baselines, input frames are sampled for each video such that the total token count roughly aligns with the number of tokens output by Veo3-1P. The SigLIP-so400m/14 model is from the PaliGemma series ((Beyer et al., 2024; Steiner et al., 2024)) which operates at 896×896 resolution with pre-pooling giving us 256 tokens per frame, and the SigLIP2-B/16 model ((Tschannen et al., 2025)) operates at 512×512 resolution without pre-pooling and outputs 1024 tokens per frame.

Learning rate follows the cosine decay schedule with linear warm-up over $1,000$ iterations. The peak learning rate is 5×10^5. The Gemma2 weights are initially frozen and gradually thawed with a linear schedule going from 0 to 0.01 over $10k$ iterations. It plateaus after $10,000$ steps so that the LLM is still updated very slowly to avoid catastrophic forgetting. We use the ADAM optimizer (Kingma & Ba, 2014) with batch size of 64 for all experiments in this subsection.

Models are trained for $32,000$ iterations in total. In cases where over-fitting was observed, early stopping at $12,000$ or $15,000$ iterations was applied. This was especially useful on the SSv2 and VATEX datasets when using a single noise injection level as training data (no 'Noise Aug.').

The MS-COCO experiments use the Karpathy splits for training, validation and test with $5,000$ test images (Karpathy & Fei-Fei, 2015).