A UNIFIED EVALUATION FRAMEWORK FOR FROZEN VISUAL MODELS ON FORECASTING TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting future events is a fundamental capability for general-purpose systems that plan or act across different levels of abstraction. Yet, evaluating whether a forecast is "correct" remains challenging due to the inherent uncertainty of the future. We propose a unified evaluation framework for assessing the forecasting capabilities of frozen vision backbones across diverse tasks and abstraction levels. Rather than focusing on single time steps, our framework evaluates entire trajectories and incorporates distributional metrics that better capture the multimodal nature of future outcomes. Given a frozen vision model, we train latent diffusion models to forecast future features directly in its representation space, which are then decoded via lightweight, task-specific readouts. This enables consistent evaluation across a suite of diverse tasks while isolating the forecasting capacity of the backbone itself. We apply our framework to nine diverse vision models, spanning image and video pretraining, contrastive and generative objectives, and with or without language supervision, and evaluate them on four forecasting tasks, from low-level pixel predictions to high-level object motion. We find that forecasting performance strongly correlates with perceptual quality and that the forecasting abilities of video synthesis models are comparable or exceed those pretrained in masking regimes across all levels of abstraction. However, language supervision does not consistently improve forecasting. Notably, video-pretrained models consistently outperform image-based ones.

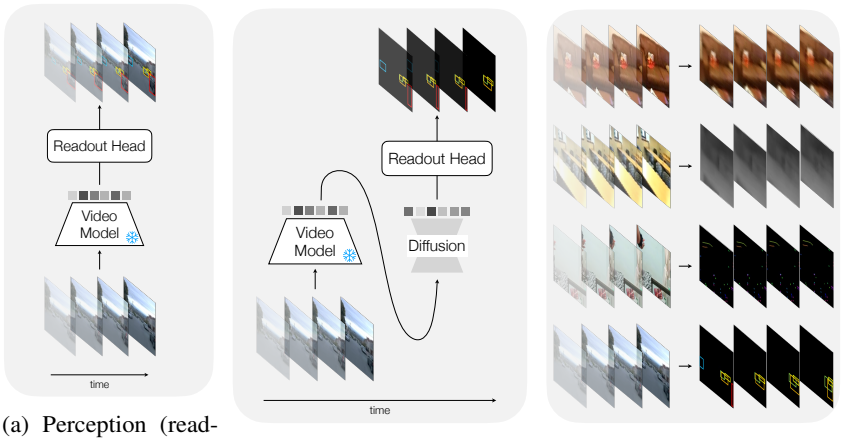
1 INTRODUCTION

The ability to see does not just reveal the present. It lets us anticipate the future and plan or act in the world accordingly. This capacity for visual forecasting is as critical for a gazelle dodging predators on the savanna as it is for a self-driving car navigating the urban jungle. At the same time, in most practical scenarios, the future is hard to predict. At any moment, countless possibilities lie ahead, and any model of the future must grapple with this inherent uncertainty.

While modern computer vision models learn representations general enough to work across multiple levels of abstraction such as DINOv2 Oquab et al. (2023) and 4DS Carreira et al. (2024), most focus on *perception* — tasks grounded in past and present frames with little to no stochasticity. Several methods of evaluations have been developed for self-supervised perception tasks, such as linear readouts, nearest-neighbors, cross-attention based, etc. However, evaluating vision models on these tasks tells us whether they understand what has already happened but may not reveal how well they can forecast what is to come.

In this paper, we shift the focus from evaluating the video *perception* capabilities of vision models to evaluating their video *forecasting* capabilities—assessing learned visual representations that can be used to predict future states of the world under uncertainty and across multiple levels of abstraction, such that diverse perceptually relevant quantities can be decoded from a single predictive representation.

We propose a unified forecasting evaluation framework built around a diffusion-based forecasting model, enabling forecasting across a range of frozen base video models and tasks: pixels, depth, point tracks, and bounding boxes. While recent work has explored using generative models for perception tasks (*e.g.*,Zhao et al. (2023); Luo et al. (2023); Li et al. (2023); Hedlin et al. (2023);



(a) Perception (readout train) (b) Forecasting (c) Downstream tasks

Figure 1: **Diffusion-based forecasting evaluation framework of frozen vision model backbones.** **(a)** *Perception-style readouts:* we train readout heads on frozen representations to perform downstream perception tasks like object detection on observed frames as in Carreira et al. (2024). We extend this setup to forecasting as follows. **(b)** *Forecasting framework:* We introduce a forecasting diffusion model that predicts future representations conditioned on frozen observed context representations. Pretrained readouts then decode these into future downstream abstractions, such as bounding boxes. **(c)** *Forecasting across abstraction levels:* We apply our approach to evaluate forecasting on tasks spanning low to high-level structure—pixels, depth, point tracks, and object detections. Each example shows 4 observed frames and a sample from the 12 forecast frames (frames 7, 10, 13, and 16). *Our results show that frozen video representations can generalize to forecasting across a wide range of downstream tasks.*

Zhang et al. (2023); Bhattad et al. (2023); Xu et al. (2025)), the reverse—using perception models for forecasting—has been less explored. We fill this gap by showing that frozen video models trained for perception *can* be effectively repurposed for forecasting, and we establish a benchmark and strong baselines to support future explorations.

Forecasting in video presents three key challenges. First, the future is inherently stochastic—multiple plausible outcomes can unfold from the same past. Second, forecasting is not about reaching a single endpoint, but about modeling how the future evolves over time as a continuous trajectory. Third, the future manifests at multiple semantic levels, from low-level pixels to mid-level motion tracks and high-level object abstractions. Our diffusion-based approach addresses all three: it captures uncertainty by generating diverse samples, models full temporal trajectories rather than single future states, and enables forecasting across a range of prediction targets, including pixels, depth, point tracks, and bounding boxes.

We extend frozen, pretrained state-of-the-art video models to forecasting tasks in two stages. First, we fit a lightweight attention-based readout head to each model for each downstream task, following the perception-based paradigm of Carreira et al. (2024) (Figure 1a). This readout head maps from the space of frozen video representations to task outputs (*e.g.* point tracks), trained using standard perception-based supervision. Then, we train a diffusion model to *forecast* future trajectories directly in the space of the frozen video representations (Figure 1b). During evaluation, we pass forecast trajectories through the readout head and assess their quality in the space of the downstream task (Figure 1c) using a suite of metrics that measure both realism and diversity—capturing the full dynamics and stochastic nature of the predicted futures.

Our diffusion-based forecasting framework enables direct, apples-to-apples comparison between perception- and synthesis-based models across all levels of visual abstraction. Our large-scale study reveals several key insights. First, forecasting ability is generally correlated with perception performance, but this association starts to break down with the strongest models. Second, video synthesis models like WALT Gupta et al. (2024) meet or exceed the forecasting performance of similarly-sized models trained with mask-based objectives when evaluated with a distribution-sensitive metric. Third, language supervision does not help forecasting performance. Fourth, models

trained solely on static images consistently underperform, highlighting the importance of temporal context in learning generalizable video representations.

2 Related Work

Video (pixel) synthesis. Early video prediction models used recurrent architectures to directly model pixel intensities Ranzato et al. (2014); Oprea et al. (2022), but struggled with long-term dynamics. Probabilistic models like SVG-LP Denton and Fergus (2018) and GANs Clark et al. (2019); Tulyakov et al. (2018); Wang et al. (2020) improved visual quality by factoring content and motion. More recently, diffusion models Ho et al. (2020) have dominated video synthesis, with models like Sora OpenAI (2024), MovieGen Polyak et al. (2025), VideoPoet Kondratyuk et al. (2023) and WALT Gupta et al. (2024) leading the way. These models capture temporal dynamics through stochastic differential equations. While some diffusion-based works address video prediction Gu et al. (2023); Xing et al. (2024); Höppe et al. (2022); Ye and Bilodeau (2024); Yang et al. (2023), they are either language-guided or operate in implicit representation spaces. We leverage a diffusion model as a general forecasting engine across a range of visual abstractions.

Task-specific video-based forecasting. Directly forecasting future pixels often fails to produce representations useful across abstraction levels. Luc et al. (2017) showed that forecasting semantic segmentation maps outperforms segmenting predicted RGB frames. They later proposed forecasting in the feature space of Mask R-CNN Luc et al. (2018), an approach similar to ours. Similarly, Vondrick et al. (2016) forecast features of AlexNet to predict actions and objects in the future. Similarly, other previous works Saric et al. (2020); Lin et al. (2021) also focus specifically on the task of forecasting segments or pixel interpolation Argaw and Kweon (2022). However, we generalize this framework: rather than relying on task-specific networks, we forecast in the frozen representation space of large pretrained video models that support a broad range of downstream tasks.

A separate line of work focuses on learning temporal dynamics from scratch using generative models or Neural Differential Equations (NDEs), such as Trajectory Flow Matching Zhang et al. (2024) and ImageFlowNet Liu et al. (2025). While these methods construct explicit, task-specific dynamical models, we ask how well general-purpose frozen video representations capture implicit dynamics that can be leveraged for forecasting via a separate diffusion-based module.

Multi-task forecasting with frozen video representations. Instead of designing task-specific forecasting models, several works have explored forecasting directly in the frozen representation space of large pretrained video models, leveraging their generality across downstream tasks. In this setup, future representations are predicted by learning lightweight forecasting heads on top of frozen features. Variants of this approach appear in recent work Rajasegaran et al. (2025); Karypidis et al. (2024). Rajasegaran et al. (2025) pretrain autoregressive models on large-scale video data and evaluate the resulting frozen features using probing tasks such as short-term interaction anticipation.

Most related to our work, DINO-Foresight Karypidis et al. (2024) and Back to the Features Baldassarre et al. (2025) forecast frozen DINOv2 Oquab et al. (2023) features using a masked transformer and autoregressive model, respectively, evaluating downstream tasks such as segmentation, depth, and surface normals at a single future time point. Unlike our approach, these works assume deterministic futures and perform single-step prediction, while we model uncertainty and evaluate the entire distribution of future trajectories.

Stochastic approaches to forecasting. In many contexts, visual forecasting is an inherently stochastic problem. A simple deterministic regressor may not necessary capture the full diversity of possible future outcomes. Some previous approaches have attempted to address this stochasticity. For example, Bhattacharyya et al. (2019) proposes a Bayesian model that jointly captures epistemic and observation aleatoric of future states. Makansi et al. (2020) uses mixture density networks to estimate the location of objects like pedestrians and vehicles from an egocentric view. In this paper, we use a diffusion model to directly learn the continuous distribution of future features.

3 Method

While prior works have explored forecasting directly in pixel space, we hypothesize that latent spaces should be better because they make the scene structure more clear and remove non-semantic, hard-

to-predict details, which should make prediction easier. Therefore, we start with frozen pretrained models, and use them to both represent the conditioning (past) video and the future video that we wish to predict. We develop a two-stage forecasting evaluation framework built around a diffusion-based forecasting module that operates directly in the space of frozen video representations. This setup allows us to extend representations trained for perception or pixel synthesis to forecasting tasks without fine-tuning. We first train lightweight readout heads to decode task-specific outputs from frozen representations. Then, we train a diffusion model to forecast future latent trajectories in the space of the frozen video representations. These forecast representations are passed through the same readouts, enabling the evaluation of nondeterministic futures across multiple semantic levels. The full pipeline is illustrated in Figure 1.

3.1 Latent Forecasting via Diffusion

We forecast future representations using a conditional denoising diffusion model Ho et al. (2020). Given a sequence of frozen representations up to time t, the diffusion model generates future latent trajectories for times $t+1 \dots T$ conditioned on the past (Figure 1b). Unlike pixel-space synthesis, our model operates in the latent space of each frozen backbone, making it architecture-agnostic and capable of comparing perception and synthesis models under the same framework.

We train one diffusion forecasting module per frozen video model under consideration. In our experiments, we condition the diffusion model on the latent encodings of $t = 4$ past frames after applying layer normalization, as it was found that the diffusion model would occasionally struggle to forecast in unnormalized latent space. We model latent encodings of a $T = 16$ frame clip, which includes the 4 past frames and 12 future frames ($16 \times 224 \times 224 \times 3$ clip). The diffusion models the latent encodings of all these frames jointly in time. Even though the diffusion model may already have information on the first t frames for conditioning, the entire clip is modeled jointly to account for temporally-entangled features.

The diffusion forecasting model never takes as conditioning any direct pixel information, only latent encodings from a given video model. This pipeline is the same whether the encoder is a video or image model. In the image encoder case, the latents are the stacked result of the image encoder on all frames.

3.2 Task Readout Heads

We decode the sampled latent trajectories using the previously trained readout heads to evaluate forecast futures. This allows us to assess prediction quality across different abstraction levels using task-appropriate metrics (Figure 1c). We train a lightweight attention-based readout head to decode the output of each frozen model into a task-specific prediction space. We focus on four tasks that vary in their level of abstraction. These are pixels, depth, point tracks, and bounding boxes. The readout heads for the four tasks follow those of Carreira et al. (2024) with the architecture of the depth readout head also used (but trained separately) for the pixel readout task (Figure 1a). These readouts are trained using standard supervised losses and provide a shared interface for comparing models across different architectures and pretraining paradigms. Importantly, readout heads are trained only on observed (past and present) frames and remain fixed during forecasting. We train a readout head for each frozen video model and downstream task pair. During inference, the transformer-based readout head processes features from all frames, conditional and forecast, simultaneously via full attention. The loss from the readout heads is not backpropagated to the diffusion model.

3.3 Evaluation Metrics

We measure performance by evaluating the accuracy and realism of the entire future trajectory rather than a static single target timestep in the future. To do so, we use two perspectives to evaluation. The first is to measure performance on a *per example* basis, measuring the statistics (mean, variance, max, min) of task-specific metrics over a set of samples for each example. Because this uses task specific metrics, it gives a reference point versus the corresponding perception task. The second is on the *dataset* level, using Fréchet Distance and variance of samples from the ground truth dataset. We argue that these metrics best consider the fact that future forecasting is inherently a stochastic task.

Per Example Metrics. For each example, we take 10 samples from our diffusion model and report the statistics of task-specific metrics from the ground truth over each sample. For pixel prediction, we use PSNR. For depth prediction, mean absolute relative error. For point tracks, we use Jaccard Distance. For box tracking, we use intersection over union. In order to account for the stochastic nature of forecasting, we report mean, minimum, and maximum of per-example samples for the given metric. For relative comparison, we also report the perception, or standalone performance on the ground truth latents on all frames, of the readout heads alongside the per example metrics.

Fréchet Distance. In order to capture the inherent stochasticity of forecasting the future, we must allow for a *distribution* over possible future trajectories and ensure that this forecast distribution is similar to that of the target ground truth data. We therefore compute the Fréchet Distance (FD) Fréchet (1957), a distribution distance metric comparing the forecast versus the ground truth set distribution over trajectories, *in the output representation of each task*. While Ng et al. (2022) employed FD for motion forecasting evaluation and Thakkar et al. (2025) for self-driving cars, action, and object interactions, we extend the use of FD for evaluating forecasts of pixels, depth, point tracks, and object bounding box tracks. Specifically, we first represent forecasts as points in some fixed dimensional space. For point tracks and box tracks, we represent each trajectory as, respectively, a vector in a 24-dimensional (2D coordinates over 12 future frames) and 48-dimensional (4 coordinates over 12 frames) spaces. For depth and pixels, we downsample each of the 12 output frames to 14×14 patches, leading to a 2352-dimensional representation space. We then fit multivariate Gaussians to the predicted and ground truth distributions in this space, and compute the Fréchet distance Dowson and Landau (1982) between them. To ensure our output is always of a fixed size, we filter out trajectories that do not contain all the available data points (*e.g.* point tracks that are not visible across all target frames due to occlusion). Note that while Heusel et al. (2017); Unterthiner et al. (2019) compute FD in the Inception embedding space, there is no Inception here. We compute FD directly in the output representation of each downstream task. Explicit details are provided in Appendix **??**.

Variance. While FD considers the realism and stochasticity of the forecast futures, it is prudent to pair it with a measure of the variance over these futures to ensure that the forecast futures are as diverse as the ground truth ones. We report the variance of the trajectories over the temporal axis, averaged over all other dimensions. This specifically assesses whether methods always forecast static future trajectories, which may be realistic but certainly not diverse.

4 Experimental Setup

4.1 Downstream Tasks and Datasets

We center our evaluation on downstream forecasting tasks designed to span multiple levels of semantic abstraction, from raw pixels to high-level object bounding boxes. This diversity allows us to probe how well frozen video representations support different kinds of future prediction and identify where video models generalize, and where they fail. We visualize each of these tasks in Figure 1(c).

Pixels. We evaluate models on the task of forecasting future RGB frames in ScanNet Dai et al. (2017). While forecasting in pixel space is highly challenging and high-dimensional, it tests low-level generative fidelity and temporal coherence. Pixel forecasting captures fine-grained dynamics but is often sensitive to misalignment or visual ambiguity. We measure pixel accuracy using PSNR.

Depth. Predicting future depth maps tests a model's ability to reason about 3D scene geometry over time. It requires some abstraction beyond raw pixels while still relying on relatively dense spatial information. Depth forecasting is particularly useful for studying how models encode physical structure and motion. We measure mean absolute relative error in ScanNet.

Point Tracks. Forecasting the trajectories of dense visual features or tracked keypoints in the Perception Test dataset Pătrăucean et al. (2023). Point tracks offer a structured yet fine-grained measure of temporal consistency and motion understanding. Because the same points persist over time, they provide a strong signal for evaluating both representation quality and future modeling. We report Average Jaccard (Doersch et al. (2022)).

Object Bounding Boxes. Forecasting future object locations as bounding boxes focuses on semantic-level understanding of object motion and interaction. This task tests whether representations capture object permanence, affordance, and dynamics—crucial for robotics or autonomous driving appli-

cations. We report Mean Intersection over Union (IoU) on the Open Waymo dataset Sun et al. (2020).

4.2 Benchmarked Models

We benchmark a set of the highest performing and largest image and video models available. See the supplementary material for the model and pretrainig specs for all models under consideration.

Image models. We benchmark SigLIP-2B Zhai et al. (2023), a 2B-parameter vision transformer trained on image–text pairs using a contrastive binary classification objective, and DINOv2 Oquab et al. (2023), a 303M-parameter vision transformer trained purely on images using a self-distillation loss without any language supervision. Since these models are not natively trained on video, we follow Carreira et al. (2024) and append learnable temporal positional embeddings to their output features. This modification enables fair comparison with video models by allowing the readout heads to exploit temporal structure when trained on top of the frozen embeddings.

Video models. We evaluate two categories of video models. The first group consists of models trained using masking-based self-supervised objectives. VideoMAE Tong et al. (2022), VideoMAEv2 Wang et al. (2023), and 4DS-e Carreira et al. (2024) are trained to reconstruct masked pixels, while V-JEPA Bardes et al. (2024) uses a feature reconstruction loss based on predictions from a teacher network. VideoPrism Zhao et al. (2024) incorporates language supervision through contrastive learning between video and text during pretraining, followed by a second stage that applies a masked reconstruction loss on video. The second group includes WALT Gupta et al. (2024), a video synthesis model trained jointly for frame prediction by conditioning on a past-frames-based signal with a probability $p_{\text{fp}} = 0.1$. We leverage this built-in capability in a pipeline referred to as Native WALT (N-WALT). N-WALT is not a new model; it is simply the pretrained WALT model used exclusively in its forecasting mode. For this pipeline, a single forward pass is performed with the past-frames-based conditioning signal to extract predictive features from the model intermediate layers. These features are then directly decoded by lightweight readout heads to produce task-specific outputs, thereby obviating the need for a separate diffusion model. We use the same layers for feature extraction as in Vélez et al. (2025).

4.3 Implementation Details

Forecasting diffusion model and readout head. Our diffusion implementation uses DDIM Song et al. (2021) and incorporates a cosine schedule Nichol and Dhariwal (2021). The underlying backbone denoiser is a vanilla 5-layer transformer. Each transformer layer employs multi-headed attention with 8 heads, utilizing 1024 total dimensions for queries, keys, and values, alongside a 2048-dimension hidden layer for the MLP. The training objective for the diffusion model minimizes the mean squared error between the denoised output of the model and the original latents, computed from a given video encoder. The diffusion model architecture is consistent across all of the underlying video models.

The training methodology for the task-specific readout heads is the same as in Carreira et al. (2024). Readout heads are attention based and trained with L2 error for pixels and depth. For point tracks, a weighted sum of Huber loss of positions and cross entropy over visibility and uncertainty is used. For box tracking, L2 loss between the labeled box coordinates and predicted position is used.

Because the base video model latents are frozen, the forecasting diffusion model and the readout head can be trained simultaneously. We found that layer normalization Ba et al. (2016) of the frozen latents is extremely important for forecasting performance. We train for 40k iterations at a batch size of 32 or an equivalent 160k iterations for a batch size of 4 for the memory intensive SigLIP. Aggregating across all experiments, we utilize approximately 144 days worth of tpu-v5 and v6 chips.

Evaluation protocol. For pixels and depth, we use the standard train and validation split on ScanNet Dai et al. (2017). For point tracks, we train in the Kubric movie dataset Greff et al. (2022) and test on the Perception Test dataset Pătrăucean et al. (2023). On box tracking, we use the train and validation split of the Waymo Open Box dataset Sun et al. (2020). For all forecasting tasks we take as context 4 frames and forecast the next 12 frames in all experiments. We sample the diffusion model 10 times per example during evaluation.

Model	Pixels			Depth			Point Tracks			Box Tracks		
	Mean ↑	Best ↑	FD ↓	Mean ↓	Best ↓	FD ↓	Mean ↑	Best ↑	FD ↓	Mean ↑	Best ↑	FD ↓
4DS-e Reg.	18.96	18.96	46.61	0.188	0.188	694.18	0.59	0.59	0.00070	0.58	0.58	2.26
4DS-e	19.89	22.03	30.95	0.1937	0.096	533.0	0.58	0.61	0.00068	0.56	0.66	1.87
WALT Reg.	21.69	21.69	14.4	0.2196	0.2196	209.86	0.61	0.61	0.00140	0.54	0.54	2.44
WALT 500M	20.4	22.55	5.46	0.230	0.138	210.1	0.64	0.68	0.00134	0.50	0.58	2.47

Table 1: **A deterministic regressor may be optimal in predicting the mean outcome, but it fails to account for the variance in possible outcomes.** Comparison of forecasting with a deterministic regression model versus a stochastic diffusion model conditioned on 4 frames. Mean and Best represents the mean and best out of 10 samples (for diffusion) on the task specific metric. For regression, there is only 1 deterministic output. For pixels, this is PSNR. For depth, Mean Absolute Relative Error. For point tracks, it is Jaccard Distance. For Box Tracks, it's IoU. FD is Frechet Distance in the output space.

Task (Dataset)	Pixels (ScanNet)		Depth (ScanNet)		Points (Perc. Test)		Boxes (Waymo)	
	FD↓	Var.(10^{-3})	FD↓	Var.(10^{-3})	FD↓(10^{-3})	Var.	FD↓	Var.
GT		*12.00*		*193*		*0.039*		*0.0032*
DINOv2	62.97	3.1	588.36	8.9	1.9	0.038	3.08	0.044
SigLIP	203.32	0.5	849.18	2.4	3.0	0.038	3.22	0.05
VideoPrism	70.30	6	882.11	7.3	0.8	0.039	2.72	0.045
VJEPA	37.08	5.1	558.28	7.7	0.63	0.039	2.85	0.048
VideoMAE	28.14	7	547.40	9.2	0.55	0.039	2.62	0.045
VideoMAEv2	33.75	5.9	578.51	7.7	0.74	0.039	2.92	0.048
4DS-h	28.29	10.0	555.11	8.5	7.88	0.039	3.24	0.054
4DS-e	30.95	6.2	533.00	11.0	0.68	0.039	1.87	0.036
WALT 500M	5.46	6.9	210.10	5.6	1.34	0.039	2.47	0.040
N-WALT 500M	6.63	5.4	217.8	4.3	1.39	0.038	3.23	0.055

Table 2: **Distributional alignment of forecast futures.** We report Fréchet Distance (lower is better) and the variance of fitted Gaussian distributions for each metric, comparing the ground truth distribution to the model's sampled forecasts. Ideally, the forecast variance should closely match the ground truth. Results show that *stronger perception models produce forecasts with distributions more aligned to the data*, reinforcing trends observed in the per-example metrics (Fig. 2).

WALT setup. We utilized WALT, a text-to-video diffusion model, as a frozen encoder, probing its intermediate layers in a setup similar to Vélez et al. (2025). WALT is designed to process 17 video frames, tokenizing them into five latent representations: one for the initial frame and four for the subsequent 16 frames. To maintain consistency with other models in this study, we sampled 16 frames, duplicated the first frame, and simulated the forward diffusion process by adding noise at timestep t. Instead of the complete multi-step generative process, the visual representation is obtained by a single forward pass through the denoiser, utilizing a null text embedding. During the single pass, the intermediate representations are extracted, discarding the initial latent representation. We use the same layers for feature extraction as in Vélez et al. (2025).

5 RESULTS

The need for a stochastic evaluation. To demonstrate the need for modeling the stochasticity of future events, we perform an ablation of our proposed diffusion forecasting module and compare it to regression-based forecasting in Table 1. We find that while regression optimizes the traditional mean-based metric, this does not account for the inherent stochasticity of forecasting. When considering metrics such as Fréchet Distance or Best-of-N that take this stochasticity into account, we find that diffusion models generally outperform the regression baselines in most cases.

Forecasting mostly correlates with perception. Overall, we observe a strong correlation between per-example metrics and the perception performance in Figure 2 where forecasting performance is displayed alongside perception performance. A more nuanced picture emerges when looking at

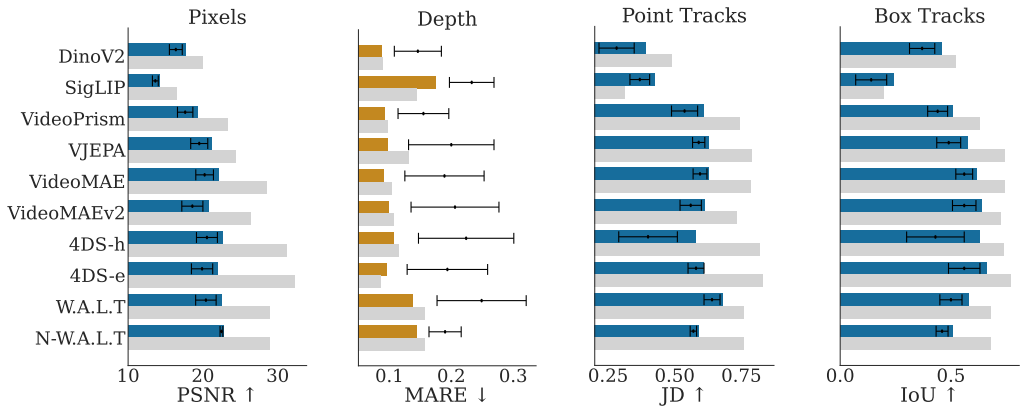


Figure 2: **Forecasting per-example metric results.** We evaluate forecasting on pixels (PSNR), point tracks (Jaccard Distance), bounding box tracks (IoU), and depth maps (Mean Absolute Relative Error) using 10 samples per example. The colored bars represent the best of n metrics for a particular task. Blue represents tasks where higher performance is better, while gold represents a metric where lower is better. We also report perception performance on each task as gray bars. Given the stochastic nature of forecasting, we report the **mean** (as whisker plot with standard deviation) and **maximum/minimum** performance (colored bars) across samples. This reveals differences in sample quality not captured by the mean alone—some models exhibit similar averages but differ significantly in their best-case outputs, highlighting variation in their predictive distributions. *Overall, we observe that stronger perception models tend to yield better forecasting performance. However, with the exception of box tracking, the best model in forecasting is never the best in perception.*

the table in more detail. Here we find that, with the exception of box tracking, the best model in perception for a given task is not the best model for forecasting. At higher levels of performance, it appears that the dynamics of certain representations are inherently easier to model than others even if other representations contain more information relevant to the downstream task.

Synthesis models like WALT achieve forecasting performance on par with or better than models trained with mask-based objectives. WALT significantly excels at pixel and depth forecasting—tasks closely aligned with its training objective—as revealed by FD when evaluating against masked representation-learning models of similar size (VideoMAE and 4DS-h). WALT also outperforms 4DS-h in both point and bounding box forecasting, while performing comparably to VideoMAE in these tasks. This outcome does not align with the lower perception performance of WALT when benchmarked against these two models for depth prediction and object tracking. It is worth noting that N-WALT does not exhibit the same performance. Since it was trained with a frame prediction objective and is conditioned on past frames, it excels at pixel forecasting, effectively capturing low-level spatiotemporal dynamics. However, it underperforms in other tasks that require higher-level semantic understanding, such as point and box tracking. This performance disparity reveals a fundamental limitation of the pixel prediction objective, suggesting that the learned features are not truly generalizable.

Language supervision does not result in better forecasting. In Table 2, we find that language-augmented models like SigLIP and VideoPrism, which were trained only on perception-style tasks, lag behind.

Video backbones outperform image ones. Notably, we find that models pretrained exclusively on image-based objectives, such as DINOv2 and SigLIP, perform poorly across most tasks, reinforcing the importance of temporal supervision during pretraining, contrary to widespread belief Baldassarre et al. (2025).

Per-example vs. distribution-level metrics. To compare our two proposed metric approaches, we first observe the per-example forecasting metrics in Figure 2. Unsurprisingly N-WALT is the strongest in forecasting pixels, given it was directly trained to do so. However, for depth forecasting, DinoV2 seems to exhibit the best per-example results. WALT with the trained diffusion head, but not N-WALT, is the best on forecasting point tracks. Interestingly, VideoMAEv2 underperforms its predecessor

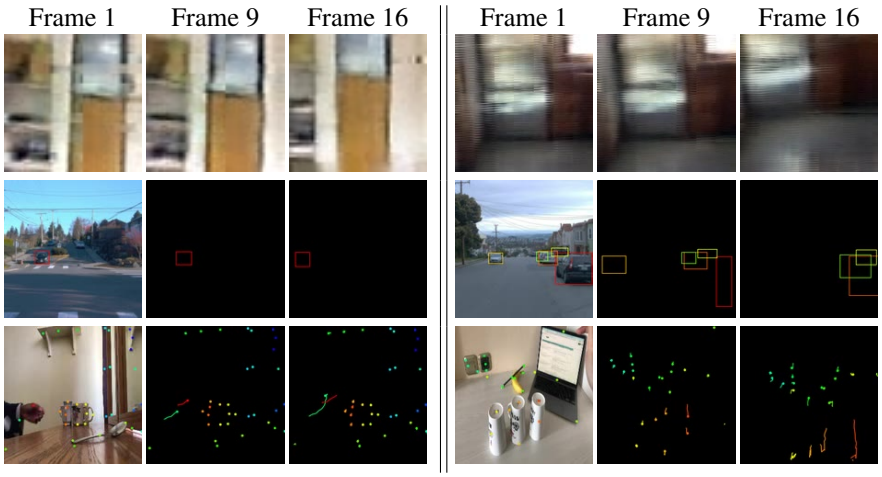


Figure 3: **Qualitative forecasts from the 4DS-e model across diverse tasks.** We condition on frames 1–4 and forecast frames 5–16. **Top:** Pixels forecasting—the model captures smooth camera motion. **Middle:** Bounding boxes—it predicts a car turning (left) and vehicle motion (right). **Bottom:** Point tracks—the model forecasts a hand rising (left) and camera motion (right). These results demonstrate that our approach generalizes well across forecasting domains and abstraction levels.

across nearly all metrics, while VideoMAE (v1) shows strong results in bounding box forecasting (per-example) and point tracks (FD), despite its smaller model size.

We next turn to the distribution level metrics in Table 2. Overall, we observe a strong correlation between the per-example metrics in Figure 2 and the distributional alignment of predicted futures with ground truth, as captured by FD and variance. However, we also find notable discrepancies emerge between the two forecasting evaluation paradigms. For instance, in depth forecasting, WALT performs strongly in terms of Frechet Distance, but DinoV2 seems to exhibit better per-example results. Similarly WALT excels on per-example metrics for point tracking, yet VideoMAE is better in terms of Frechet Distance on this task. This discrepancy highlights how small-sample evaluations can obscure poor distributional alignment.

We find that all models on most tasks, with the exception of point tracking, struggle to approach the variance of the ground truth datasets. Even though WALT performs relatively well on pixel forecasting with respect to the Frechet Distance metric, it still does not capture the full extent of the underlying variance in pixel space. The variance disparity is especially apparent with depth forecasting; this suggests that current models particularly struggle to model visual information relevant to this particular domain.

Qualitative. We visualize forecasts from the 4DS-e model in Figure 3. These results demonstrate that our approach generalizes well across multiple forecasting domains and abstraction levels.

6 Discussion

We proposed a unified evaluation framework of frozen vision backbone models in forecasting tasks. Our central findings are first, that forecasting performance in frozen pretrained video models closely tracks their perception performance up to a point. At higher levels of performance, this association starts to break down. Second, as expected, WALT, a video synthesis model explicitly trained to generate future frames, significantly outperforms masked video models on low-level forecasting tasks such as pixel and depth prediction. However, WALT's forecasting strength is mixed for mid-level structured tasks like point tracks and object bounding boxes when comparing masked models of similar size. Third, language supervision alone does not appear to improve forecasting ability, underscoring the importance of temporal visual learning for anticipating future states. Lastly, our results clearly show that video backbone models outperform their image-based counterparts in supporting future-forecasting tasks.

REFERENCES

Dawit Mureja Argaw and In So Kweon. Long-term video frame interpolation via feature propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3543–3552, 2022.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Bayesian prediction of future street scenes using synthetic likelihoods. In *International Conference on Learning Representations*, 2019.

Anand Bhattad, Daniel McKee, Derek Hoiem, and David Forsyth. Stylegan knows normal, depth, albedo, and more. *Advances in Neural Information Processing Systems*, 36:73082–73103, 2023.

João Carreira, Dilara Gokay, Michael King, Chuhan Zhang, Ignacio Rocco, Aravindh Mahendran, Thomas Albert Keck, Joseph Heyward, Skanda Koppula, Etienne Pot, et al. Scaling 4d representations. *arXiv preprint arXiv:2412.15212*, 2024.

Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International conference on machine learning*, pages 1174–1183. PMLR, 2018.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022.

DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Maurice Fréchet. Sur la distance de deux lois de probabilité. In *Annales de l'ISUP*, volume 6, pages 183–198, 1957.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897*, 2023.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *ECCV*, 2024.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems*, 36:8266–8279, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022.

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Dino-foresight: Looking into the future with dino, 2024. URL https://arxiv.org/abs/2412.11673.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023.

Zihang Lin, Jiangxin Sun, Jian-Fang Hu, Qizhi Yu, Jian-Huang Lai, and Wei-Shi Zheng. Predictive feature learning for future segmentation prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7365–7374, 2021.

Chen Liu, Ke Xu, Liangbo L Shen, Guillaume Huguet, Zilong Wang, Alexander Tong, Danilo Bzdok, Jay Stewart, Jay C Wang, Lucian V Del Priore, et al. Imageflownet: Forecasting multiscale image-level trajectories of disease progression with irregularly-sampled longitudinal medical images. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future of semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 648–657, 2017.

Pauline Luc, Camille Couprie, Yann Lecun, and Jakob Verbeek. Predicting future instance segmentation by forecasting convolutional features. In *Proceedings of the european conference on computer vision (ECCV)*, pages 584–599, 2018.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023.

Osama Makansi, Ozgun Cicek, Kevin Buchicchio, and Thomas Brox. Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4354–4363, 2020.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

OpenAI. Sora, 12 2024. URL https://openai.com/sora/.

Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2806–2826, 2022. doi: 10.1109/TPAMI.2020.3045007.

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

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet,

Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2025. URL https://arxiv.org/abs/2410.13720.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HYEGXFnPoq.

Jathushan Rajasegaran, Ilija Radosavovic, Rahul Ravishankar, Yossi Gandelsman, Christoph Feichtenhofer, and Jitendra Malik. An empirical study of autoregressive pre-training from videos. *arXiv preprint arXiv:2501.05453*, 2025.

MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014.

Josip Saric, Marin Orsic, Tonci Antunovic, Sacha Vrazic, and Sinisa Segvic. Warp to the future: Joint forecasting of features and feature motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10657, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Neerja Thakkar, Tara Sadjadpour, Jathushan Rajasegaran, Shiry Ginosar, and Jitendra Malik. Poly-autoregressive prediction for modeling interactions. In *CVPR*, 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.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. *ICLR workshop*, 2019.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 98–106, 2016.

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. *arXiv preprint arXiv:2502.07001*, 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*, pages 14549–14560, 2023.

Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1160–1169, 2020.

Zhen Xing, Qi Dai, Zejia Weng, Zuxuan Wu, and Yu-Gang Jiang. Aid: Adapting image2video diffusion models for instruction-guided video prediction. *arXiv preprint arXiv:2406.06465*, 2024.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? *ICLR*, 2025.

Siyuan Yang, Lu Zhang, Yu Liu, Zhizhuo Jiang, and You He. Video diffusion models with local-global context guidance. *arXiv preprint arXiv:2306.02562*, 2023.

Xi Ye and Guillaume-Alexandre Bilodeau. Stdiff: Spatio-temporal diffusion for continuous stochastic video prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6666–6674, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.

Xi Nicole Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. *Advances in Neural Information Processing Systems*, 37:107198–107224, 2024.

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*, 2024.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023.