



Figure 4: **Forecasting performance correlates with perceptual ability.** We evaluate forecasting on **pixels**, **point tracks**, **bounding boxes**, and **depth** using 10 samples per example and report the normalized max (min for lower-is-better metrics) performance per task. We compare with models’ performance on the perception-style variant of each task on observed frames. Data points correspond to model types. *In general, we find a linear correlation between forecasting and perception performance on these short-horizon tasks. However, for the best performing models this relationship is somewhat more complex.*

## A APPENDIX

**Limitations.** Our forecasting evaluation framework has a few key limitations. First, the datasets studied, while diverse, often lack complex or ambiguous motion, limiting the generality of the tasks. Second, the introduction of a diffusion model is a potential confounding variable in evaluating the forecasting capabilities of frozen models. To mitigate this risk, we opt to use a simple “vanilla” transformer to provide a standardized and simple forecasting module while preventing an overly powerful, task-specific forecaster from compensating for weaknesses in a backbone’s representations. We note that while diffusion provides a unified and expressive forecasting mechanism, it is computationally expensive and sensitive to sample count. Third, forecasting on frozen perception features may impose representational mismatches, especially for tasks requiring fine-grained temporal dynamics. Lastly, we use video models with 16-frame contexts, which restricts our ability to assess long-horizon forecasting. That said, our evaluation framework is generic in that it will apply in exactly the same way to longer-context models, once they become available.

### A.1 DETAILS ON THE FRÉCHET DISTANCE COMPUTATION

Let  $S_g = \{g_1, g_2, \dots, g_n\}$  be the set of  $n$  ground truth trajectories and  $S_p = \{p_1, p_2, \dots, p_m\}$  be the set of  $m$  predicted trajectories. We first represent each trajectory as a vector in a  $d$ -dimensional space, where the dimensionality depends on the specific forecasting task:

1. **Point tracks.** Each trajectory consists of the 2D coordinates of a point over 12 future frames. By concatenating these coordinates, we form a vector  $v_i \in \mathbb{R}^{24}$ .
2. **Box tracks.** Each trajectory represents the 4 coordinates of an axis-aligned 2D bounding box over 12 future frames. This results in a vector  $v_i \in \mathbb{R}^{48}$ .
3. **Depth and pixel predictions.** Each trajectory is a sequence of 12 dense prediction frames. We downsample each frame to  $14 \times 14$  and concatenate them, yielding a vector  $v_i \in \mathbb{R}^{2352}$ .

For a given task, all ground truth and predicted trajectories are thus represented as vectors in the same space  $\mathbb{R}^d$ .

After collecting the ground truth and predicted trajectories, we model their distributions by fitting a multivariate Gaussian to each set. We then model the distribution of these vectors by fitting a

| Model                         | Type  | Variant               | Resolution | Params | Training Time |
|-------------------------------|-------|-----------------------|------------|--------|---------------|
| DINOv2 Oquab et al. (2023)    | I     | DINOv2_L/14           | 224        | 303M   | 637.5k steps  |
| SigLIP Zhai et al. (2023)     | I/L   | PaLI3-SigLIP-G-opt/14 | 224        | 2B     | 107k steps    |
| VideoPrism Zhao et al. (2024) | V/M/L | v_giant               | 288        | 1.1B   | 500k steps    |
| VJEPa Bardes et al. (2024)    | V/M   | L                     | 224        | 300M   | 90k steps     |
| VideoMAE Tong et al. (2022)   | V/M   | H                     | 224        | 600M   | 1600 epochs   |
| VideoMAEv2 Wang et al. (2023) | V/M   | g                     | 224        | 1B     | 1200 epochs   |
| 4DS-h Carreira et al. (2024)  | V/M   | h                     | 224        | 639M   | 488,282 steps |
| 4DS-e Carreira et al. (2024)  | V/M   | e                     | 224        | 4B     | 488,282 steps |
| WALT Gupta et al. (2024)      | V/S/L | 419M                  | 128        | 419M   | 528k steps    |

Table 3: **Pretraining specs for the frozen models under consideration.** We benchmark a wide array of the strongest (I)mage and (V)ideo models trained in a (M)asking or a (S)ynthesis paradigm with or without the use of (L)anguage.

multivariate Gaussian to each set. Specifically, we estimate the parameters of the ground truth distribution  $\mathcal{N}_g(\mu_g, \Sigma_g)$  using the sample mean and unbiased sample covariance from the set  $S_g$ :

$$\mu_g = \frac{1}{n} \sum_{i=1}^n g_i$$

$$\Sigma_g = \frac{1}{n-1} \sum_{i=1}^n (g_i - \mu_g)(g_i - \mu_g)^T$$

The parameters for the predicted distribution,  $\mu_p$  and  $\Sigma_p$  of  $\mathcal{N}_p(\mu_p, \Sigma_p)$ , are computed similarly from the set of predicted trajectories  $S_p$ .

The Fréchet Distance (FD) is then defined as the squared Wasserstein-2 distance between these two estimated Gaussian distributions, and is calculated using the formula from Dowson and Landau (1982):

$$FD^2 = \|\mu_g - \mu_p\|_2^2 + \text{Tr}(\Sigma_g + \Sigma_p - 2(\Sigma_g \Sigma_p)^{1/2})$$

where  $\|\cdot\|_2^2$  denotes the squared Euclidean norm and  $\text{Tr}(\cdot)$  is the trace of a matrix. The term  $(\Sigma_g \Sigma_p)^{1/2}$  represents the matrix square root of the product of the two covariance matrices. This metric provides a single scalar value that quantifies the dissimilarity between the distribution of predicted trajectories and that of the ground truth.

## A.2 PRETRAINING SPECS FOR BACKBONE MODELS

Table 3 lists the pretraining specs for all the vision models under consideration.

## A.3 RESULTS APPENDIX

We include the full numerical results underlying Figure 2 in Table 5.

## B BROADER IMPACTS

This paper focuses on evaluating the performance of video models on forecasting. While it is true that the potential applications of video models in general do have a multitude of wide social ramifications such as surveillance or generation of disinformation, the immediate real world impact of the results in this paper are very limited. They are difficult to precisely predict over the long run. The topics explored in the paper fall under the realm of foundational research. The timescale of forecasting is no more than 3 seconds, and none of the results are state-of-the-art or have immediate applications.

Table 4: Model Performance Metrics

| Model       | Pixel PSNR $\uparrow$ |        |       |            | Depth Absolute Relative Error $\downarrow$ |       |       |            | Jaccard Distance $\uparrow$ |       |      |            | IoU $\uparrow$ |       |      |            |
|-------------|-----------------------|--------|-------|------------|--|-------|-------|------------|-----------------------------|-------|------|------------|----------------|-------|------|------------|
|             | Mean                  | Std    | Max   | Perception | Mean                                       | Std   | Min   | Perception | Mean                        | Std   | Max  | Perception | Mean           | Std   | Max  | Perception |
| DINOv2      | 16.38                 | 0.857  | 17.78 | 19.97      | 0.146                                      | 0.038 | 0.088 | 0.089      | 0.28                        | 0.066 | 0.39 | 0.49       | 0.37           | 0.057 | 0.46 | 0.52       |
| SigLIP      | 13.64                 | 0.39   | 14.29 | 16.48      | 0.233                                      | 0.036 | 0.175 | 0.144      | 0.37                        | 0.037 | 0.42 | 0.31       | 0.14           | 0.070 | 0.24 | 0.20       |
| VideoPrism  | 17.63                 | 1.03   | 19.32 | 23.33      | 0.155                                      | 0.041 | 0.093 | 0.098      | 0.54                        | 0.049 | 0.61 | 0.74       | 0.44           | 0.045 | 0.51 | 0.63       |
| VJEPa       | 19.51                 | 1.14   | 21.26 | 24.41      | 0.2  | 0.069 | 0.098 | 0.132      | 0.59                        | 0.023 | 0.63 | 0.79       | 0.49           | 0.054 | 0.58 | 0.74       |
| VideoMAE    | 20.23                 | 1.1834 | 22.13 | 28.54      | 0.189                                      | 0.064 | 0.091 | 0.104      | 0.60                        | 0.026 | 0.63 | 0.79       | 0.56           | 0.038 | 0.62 | 0.74       |
| VideoMAEv2  | 18.59                 | 1.41   | 20.81 | 26.44      | 0.206                                      | 0.071 | 0.1   | 0.107      | 0.56                        | 0.040 | 0.61 | 0.73       | 0.56           | 0.053 | 0.64 | 0.72       |
| 4DS-h       | 20.54                 | 1.40   | 22.69 | 31.24      | 0.224                                      | 0.077 | 0.107 | 0.1156     | 0.40                        | 0.110 | 0.58 | 0.82       | 0.43           | 0.130 | 0.63 | 0.74       |
| 4DS-e       | 19.89                 | 1.42   | 22.03 | 32.26      | 0.1937                                     | 0.065 | 0.096 | 0.086      | 0.58                        | 0.029 | 0.61 | 0.83       | 0.56           | 0.071 | 0.66 | 0.77       |
| WALT 500M   | 20.4                  | 1.37   | 22.55 | 28.95      | 0.249                                      | 0.072 | 0.138 | 0.158      | 0.64                        | 0.030 | 0.68 | 0.76       | 0.50           | 0.050 | 0.58 | 0.68       |
| N-WALT 500M | 22.48                 | 0.19   | 22.78 | 28.95      | 0.19                                       | 0.026 | 0.145 | 0.158      | 0.57                        | 0.012 | 0.59 | 0.76       | 0.46           | 0.027 | 0.51 | 0.68       |

Table 5: Numerical table of results presented in Figure 2. For each metric, the first three columns are the statistics of 10 samples taken from a diffusion model with 4 frames of context. The fourth column is perception (no forecasting) performance for reference.

## C TIMESCALES IN DATASET EVALUATION

We utilize 3 datasets for evaluation in the paper. They are ScanNet Dai et al. (2017), Perception Test Pătrăucean et al. (2023) and the Waymo Open Box dataset Sun et al. (2020). For all of them, we use the general setup in Carreira et al. (2024). The number of frames forecast is always 12, but the effective sampling rate differs on each dataset. The range of forecasting thus varies from roughly half a second up to 3 seconds. On the Waymo dataset, the videos are sampled at 5 frames per second. On ScanNet, sampling is approximately 25 frames per second. Perception Test videos are sampled at a stride of 4.