



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

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 14×14 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 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, μ_p and Σ_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

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.

(1982):

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

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

A.3.1 FORECASTING PER-EXAMPLE METRICS

We include the full numerical results underlying Figure 2 in Table 4.

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 4: 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.

A.3.2 N-WALT METRICS ACROSS DIFFERENT NOISE LEVELS

The WALT model was jointly trained for frame prediction to enable long video generation through autoregressive prediction. To mitigate the domain shift between training, where ground-truth long videos are available, and inference, where the model predicts subsequent frames based on its own generated output, WALT incorporates noise conditioning augmentation Ho et al. (2022) during

training. This augmentation involves applying noise to the past-frames-based conditioning signal. Concretely, conditioning noise is added in accordance with a noise schedule, by sampling a noise level as $t_n \sim \mathcal{U}(0, t_{\max})$, where $t_{\max} = 300$. We evaluated the performance of N-WALT by varying the conditioning noise level during inference across the range $[0, t_{\max}]$. The performance metrics are detailed in Tables 5 and 6, where bold values denote the best task-specific metrics presented in Tables. 4 and 2, respectively.

Noise level	Pixel PSNR \uparrow			Depth Absolute Relative Error \downarrow			Jaccard Distance \uparrow			IoU \uparrow		
	Mean	Std	Max	Mean	Std	Min	Mean	Std	Max	Mean	Std	Max
0	21.46	0.310	21.95	0.237	0.032	0.186	0.500	0.020	0.533	0.407	0.039	0.471
10	22.48	0.188	22.78	0.190	0.018	0.160	0.564	0.010	0.580	0.462	0.027	0.507
20	22.38	0.199	22.70	0.189	0.020	0.156	0.566	0.010	0.582	0.455	0.030	0.504
30	22.33	0.214	22.67	0.189	0.022	0.153	0.565	0.012	0.584	0.453	0.032	0.506
40	22.37	0.230	22.74	0.187	0.024	0.149	0.571	0.012	0.590	0.454	0.034	0.510
50	22.39	0.249	22.79	0.186	0.026	0.145	0.570	0.014	0.593	0.457	0.037	0.517
100	22.19	0.331	22.73	0.188	0.036	0.132	0.562	0.020	0.593	0.447	0.048	0.524
150	21.88	0.412	22.54	0.193	0.044	0.124	0.570	0.025	0.610	0.439	0.056	0.531
200	21.59	0.488	22.37	0.194	0.053	0.113	0.553	0.031	0.602	0.433	0.062	0.534
250	21.30	0.561	22.19	0.200	0.059	0.109	0.553	0.037	0.610	0.428	0.066	0.535
300	20.94	0.635	21.95	0.205	0.066	0.104	0.550	0.041	0.611	0.425	0.070	0.539

Table 5: Per-example metrics by varying the conditioning noise level.

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.

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.

Noise level	Pixels (ScanNet)		Depth (ScanNet)		Points (Perc. Test)		Boxes (Waymo)	
	FD↓	Var.(10^{-3})	FD↓	Var.(10^{-3})	FD↓(10^{-3})	Var.	FD↓	Var.
0	8.09	4.69	250.98	1.15	1.563	0.0378	2.72	0.0427
10	6.55	5.38	223.99	3.73	1.544	0.0379	3.23	0.0548
20	6.80	5.31	222.91	4.04	1.482	0.0378	2.89	0.0472
30	6.86	5.28	221.39	4.08	1.577	0.0377	2.90	0.0465
40	6.76	5.34	222.35	3.92	1.387	0.0379	2.72	0.0432
50	6.64	5.40	217.80	4.27	1.480	0.0378	3.13	0.0489
100	6.74	5.28	219.83	4.07	1.448	0.0378	2.60	0.0412
150	7.14	5.12	229.02	3.59	1.525	0.0378	2.66	0.0416
200	7.43	4.97	220.84	4.10	1.540	0.0378	2.78	0.0465
250	7.70	4.81	228.10	3.37	1.467	0.0378	2.89	0.0431
300	8.09	4.61	226.96	3.35	1.646	0.0369	3.04	0.0506

Table 6: Distribution-level metrics by varying the conditioning noise level.

Model	Pixels (ScanNet)	Depth (ScanNet)	Points (Perc. Test)	Boxes (Waymo)
VideoMAEv1	6.35**	5.98**	2.34***	9.13†
VideoMAEv2	23.85**	23.35**	6.33‡	6.17‡
DinoV2	12.83**	12.48**	16.62†	16.55†
VideoPrism	14.90**	13.85**	3.98‡	3.86‡
SigLip	21.74**	21.56**	18.62†	8.01‡
VJEPa	31.08**	31.05**	9.19†	9.13†
4DS-H	23.25**	22.38**	5.05§	4.04§
4DS-e	45.41**	45.46**	13.14†	13.11†
Walt	45.75*	45.77*	13.71†	13.76†
Native Walt	9.78**	8.99**	9.63†	9.00**

Table 7: Experiment training time in hours. Both readout head and diffusion model were trained in parallel, but not end-to-end (stop gradient). Models were trained on TPUs. * represents 16 v5p chips, ** represents 4 v5p, *** 6 v5p, †4 v6e, ‡16 v6e, and §32 v5p.

Model	Pixels			Depth		
	Mean ↑	Best ↑	FD ↓	Mean ↓	Best ↓	FD
4DS-e 5 Samples	20.09	21.76	31.74	0.193	0.119	538.54
4DS-e 10 Samples	19.89	22.03	30.95	0.194	0.096	533.00
4DS-e 15 Samples	20.02	22.36	29.85	0.194	0.087	538.24

Table 8: Ablation of performance at different number of samples per example.

Model	Pixels				Depth			
	Perception	Mean ↑	Best ↑	FD ↓	Perception	Mean ↓	Best ↓	FD
4DS-e 512	29.99	20.40	22.39	29.74	0.087	0.187	0.096	527.72
4DS-e 1024	32.26	19.89	22.03	30.95	0.086	0.194	0.096	533.00
4DS-e 2048	31.62	20.50	22.58	30.46	0.088	0.192	0.093	530.64

Table 9: Ablation of performance at different capacities of the transformer-based readout head. 1024 is the default number used in the paper.