Representations Before Pixels: Semantics-Guided Hierarchical Video Prediction

Efstathios Karypidis[1,3] * Spyros Gidaris[2] Nikos Komodakis[1,4,5]

[1]Archimedes, Athena Research Center, Greece [2]valeo.ai
[3]National Technical University of Athens [4]University of Crete
[5]IACM-Forth

Abstract. Accurate future video prediction requires both high visual fidelity and consistent scene semantics, particularly in complex dynamic environments such as autonomous driving. We present `Re2Pix`, a hierarchical video prediction framework that decomposes forecasting into two stages: semantic representation prediction and representation-guided visual synthesis. Instead of directly predicting future RGB frames, our approach first forecasts future scene structure in the feature space of a frozen vision foundation model, and then conditions a latent diffusion model on these predicted representations to render photorealistic frames. This decomposition enables the model to focus first on scene dynamics and then on appearance generation. A key challenge arises from the train-test mismatch between ground-truth representations available during training and predicted ones used at inference. To address this, we introduce two conditioning strategies, nested dropout and mixed supervision, that improve robustness to imperfect autoregressive predictions. Experiments on challenging driving benchmarks demonstrate that the proposed semantics-first design significantly improves temporal semantic consistency, perceptual quality, and training efficiency compared to strong diffusion baselines. We provide the implementation code at https://github.com/Sta8is/Re2Pix.

1 Introduction

Video prediction plays a central role in autonomous systems, where anticipating how a scene will evolve is essential for long-horizon reasoning and decision making [15,17,21]. In driving scenarios, the ability to accurately forecast how visual scenes evolve, from the motion of vehicles and pedestrians to the subtleties of lighting and occlusions, is not merely a perceptual convenience but a prerequisite for safe planning [18,63,73]. Yet learning such predictive models from raw video requires simultaneous mastery of high-level semantics (what objects are present and how they interact) and photorealistic details (how the scene appears) across temporal scales, a challenge that remains largely unsolved.

Most modern approaches adopt an end-to-end paradigm: future frames are predicted directly in the latent space of a VAE, typically using diffusion models [26,50]. While effective, this paradigm suffers from a fundamental limitation:

* Corresponding author: `e.karypidis@athenarc.gr`

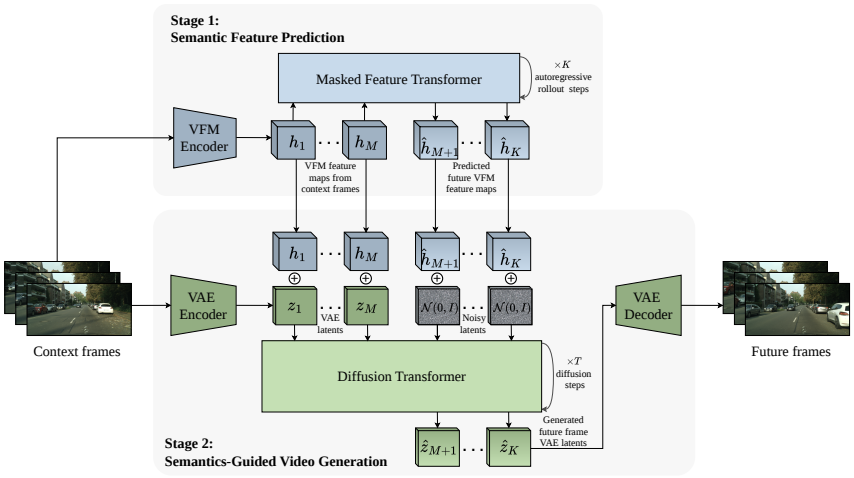

Fig. 1: Overview of the proposed `Re2Pix` hierarchical framework during inference. In Stage 1, semantic features $h_{1:M}$ of the context frames are extracted from a vision foundation model (VFM) encoder E_h and fed into a masked feature transformer to autoregressively predict (in frame-wise way) future semantic VFM features $\hat{h}_{M+1:K}$. In Stage 2, both the past $\hat{h}_{1:M}$ and predicted features $\hat{h}_{M+1:K}$ condition the diffusion transformer G_z to generate future VAE latents $\hat{z}_{M+1:K}$, which are then decoded into RGB future frames.

semantic structure and fine-grained visual details are deeply entangled within the same latent representation. Consequently, the model must simultaneously infer scene dynamics and render photorealistic appearance, often leading to temporal semantic inconsistencies such as object identity drift, structural degradation, or flickering artifacts. More fundamentally, this entanglement slows convergence, increases data requirements, and makes it difficult to reason about or control each component independently. Recent work has attempted to inject semantic structure into diffusion models by aligning intermediate features with pretrained representations via auxiliary distillation objectives [84,86]. While such alignment can guide representation learning, the diffusion model still performs forecasting and rendering within a single latent space, leaving their roles implicitly coupled.

We ask a different question: *can semantic forecasting and visual synthesis be explicitly separated, while still enabling coherent generative prediction?*

Our primary contribution is to introduce a hierarchical framework that decomposes video prediction into two interacting but distinct stages. First, we forecast future scene structure in the feature space of a frozen vision foundation model (VFM). These representations capture high-level semantics while abstracting away low-level appearance. Second, we condition a latent diffusion model on the predicted representations to render photorealistic frames. This separation yields a structured predictive pipeline: the first stage models tem-

poral dynamics in a representation space, while the second stage specializes in appearance synthesis conditioned on forecasted structure.

This clean separation, however, introduces a critical challenge. During training, the diffusion model has access to clean, ground-truth semantic features from future frames. At inference, it must instead rely on autoregressively predicted features, which inevitably accumulate errors. Naively supervising the generator solely on clean semantics creates a severe train–test mismatch: the model overfits to near-perfect conditioning signals and degrades sharply—often producing blurry or incoherent outputs—when exposed to noisy, imperfect forecasts. We demonstrate that bridging this distribution shift is essential for effective hierarchical video prediction.

Our second, technical contribution lies in modeling this bridge between semantic forecasting and generative synthesis. We adopt an early fusion strategy that token-wise merges VFM features with VAE latents at the input level, providing stable conditioning without increasing token count. To mitigate the conditioning mismatch, we introduce two complementary strategies. First, *nested dropout* stochastically truncates feature channels during training, encouraging robustness to errors in the forecasted representation. Second, *mixed supervision* exposes the generator to both ground-truth and predicted features (90/10 mixture), directly regularizing against over-reliance on idealized semantics. Together, these mechanisms enable the diffusion model to operate reliably under imperfect, autoregressively predicted features.

Concretely, our framework, `Re2Pix`, extracts semantic representations using a frozen VFM encoder such as DINOv2 [49], and trains a lightweight masked Transformer to autoregressively predict future features in this space. A latent video diffusion model, implemented with a DiT backbone [50] in the VAE latent space [69], is then conditioned directly on the *forecasted* semantic features to render future frames. In contrast to alignment-based approaches such as REPA [84] and VideoREPA [86], which encourage feature similarity between diffusion and semantic spaces, we treat forecasted VFM representations as an explicit intermediate generative variable.

Our main contributions are:

- We introduce `Re2Pix`, a hierarchical semantic-to-pixel framework that separates video forecasting into semantic representation prediction and semantics-driven visual generation. To our knowledge, we are the first to demonstrate that VFM feature prediction can effectively guide hierarchical video diffusion.
- We propose two robust conditioning strategies—*nested dropout* and *mixed supervision*—to close the train-test gap and improve robustness to imperfect, autoregressively generated features.
- Extensive results demonstrate that `Re2Pix` achieves substantial improvements over strong baselines in temporal semantic fidelity and perceptual quality, while significantly accelerating training convergence (up to 7× for generation and 14× for segmentation metrics).

2 Related Work

Video Generation and Prediction. Video prediction has evolved from autoregressive models operating directly in pixel space [14, 19, 40, 46, 66, 72] to hierarchical and structured approaches that model temporal dynamics more effectively [27, 45, 70, 77, 79]. Transformer-based architectures have further advanced the field by leveraging autoregressive or masked modeling objectives to capture long-range temporal dependencies [22, 71, 82, 83]. State-of-the-art video prediction systems typically operate in the latent space of a learned variational autoencoder [37, 55, 69], where generative models predict future latent codes rather than raw pixels [8, 18, 22, 82, 83]. Recent models trained on large-scale video data produce temporally coherent, photorealistic sequences [9, 24, 52, 80]. In controllable or world-modeling settings, Vista [18] offers a generalizable driving world model with precise spatiotemporal control, while Cosmos-Predict [1, 3, 48] and Cosmos-Transfer [2] enable multi-modal conditional generation for simulation and robotics.

Unlike these approaches, which predict future frames directly in pixel or VAE-latent space, our method introduces a hierarchical formulation that first forecasts high-level VFM semantic features and then generates pixels conditioned on them. This design improves temporal semantic consistency and reduces the burden on the diffusion model by providing stronger structural guidance. As a result, our framework bridges semantic forecasting with generative video modeling while maintaining competitive training efficiency.

Semantic Future Prediction. A line of work in future frame prediction focuses on forecasting semantic information rather than raw RGB values [32, 47, 61, 67], typically predicting representations extracted from pre-trained networks. Early approaches [30, 35, 44, 57] targeted intermediate features or outputs of task-specific scene understanding models, such as Mask-RCNN [23] and Segmenter [59].

More recently, DINO-Foresight [34] and DINO-WM [88] forecast dense, patch-wise semantic from vision foundation models like DINOv2 [49], which, thanks to large-scale pre-training, generalize effectively to diverse tasks and new scenes without retraining. DINO-WM applies this approach to world modeling and action-conditioned planning in simulated environments, whereas DINO-Foresight focuses on multi-task dense semantic forecasting in real-world driving scenarios. Subsequent works extended this paradigm with diffusion-based formulations [68] and by scaling to larger datasets and models [6].

Relatedly, V-JEPA methods [5, 7] learn visual representations by predicting masked video regions, but their objective is representation quality rather than forecasting future frames. In contrast, our work leverages VFM feature prediction as a hierarchical intermediate for RGB generation, enabling future frame synthesis with strong temporal semantic consistency while maintaining training efficiency.

Leveraging VFM features for visual generation. A growing body of work explores using features from pre-trained vision foundation models (VFMs) [49, 58, 62, 65]

as strong priors for generative modeling. One line of methods aligns VAE latent spaces with VFM features through distillation losses [11, 28, 43, 56, 81]. Another aligns intermediate diffusion features with VFM representations [41, 84], an approach pioneered by REPA [84], which substantially accelerates diffusion training and improves image generation quality. These ideas have been extended to video generation [31, 76, 86], where VFM-guided objectives improve temporal semantic consistency and 3D geometry when *fine-tuning* pretrained video diffusion transformers. However, these works do not demonstrate improved training convergence for video diffusion models trained *from scratch*, nor do they address video prediction settings.

Recent work further leverages VFMs by jointly modeling low-level VAE latents and high-level VFM features within diffusion [38, 75], or by generating only high-dimensional VFM representations that are subsequently decoded to RGB [87], both of which yield faster convergence and improved fidelity. Semantic representations have also been explored in hierarchical image synthesis pipelines that first predict global semantic representations and then generate VAE latents conditioned on them [42, 51].

In contrast to these approaches—which treat VFM features as static conditioning signals or as alternative generative latents—our method evolves VFM features through time, using them as a dynamic hierarchical intermediate for video generation. This semantics-guided forecasting enables temporally coherent and content-aware future frame synthesis, and, to our knowledge, we are the first to leverage VFM feature prediction for hierarchical video prediction.

3 Methodology

We address the video prediction task, where the goal is to forecast future frames given a sequence of past observations. Let $x = (x_1, \ldots, x_K)$ be a video sequence of K frames. Given the first M frames ($M < K$), the task is to predict the remaining $K - M$ frames. To solve this, we propose a hierarchical framework that decouples the problem into two stages (see Figure 1):

High-Level Semantic Prediction We extract high-level semantic representations of the input frames using a pretrained vision foundation model. These representations capture the essential structure of the scene while abstracting low-level details. In the first stage, the model predicts future frames in this semantic space, allowing it to focus on structural reasoning before synthesizing fine-grained visuals.

Semantics-Guided Video Generation In the second stage, a latent video diffusion model generates future frames within the compact latent space of a Variational Autoencoder (VAE), which preserves sufficient detail for high-fidelity reconstruction. The previously predicted semantic representations guide the diffusion process, ensuring the generated content remains consistent with the scene's semantic dynamics. Finally, the VAE decoder reconstructs the output frames in pixel space, producing photorealistic future frames aligned with the predicted semantics.

An overview of our `Re2Pix` framework during inference is shown in Figure 1. This two-stage approach effectively separates semantic reasoning from visual synthesis, simplifying the video prediction task. Subsection 3.1 details the semantic prediction stage, Subsection 3.2 describes the semantics-guided generation process, Subsection 3.3 presents the architecture of the diffusion model, and Subsection 3.4 outlines the training strategies that ensure robust semantic conditioning.

3.1 High-Level Semantic Prediction

In the first stage, we extract high-level semantic representations from the input frames using a pretrained Vision Foundation Model (VFM). Specifically, we employ the DINOv2 [49] image encoder $E_h(\cdot)$, which processes each frame x_t independently to produce a feature map h_t:

$$h_t = E_h(x_t), \quad t = 1, \ldots, M \tag{1}$$

Here, h_t has dimensions $H_h \times W_h \times C_h$, where $H_h \times W_h$ are the spatial dimensions and C_h is the number of feature channels. These features capture the scene's semantic structure while abstracting low-level details.

Next, we predict future features autoregressively using a feature generation model $G_h(\cdot)$. Given the context features $(h_1, \ldots, h_M)$, $G_h(\cdot)$ generates the remaining $K - M$ frames one step at a time:

$$h_{M+1}, \ldots, h_K = G_h(h_1, \ldots, h_M). \tag{2}$$

We adopt the masked transformer architecture from [34] for $G_h(\cdot)$. During training, the model takes $M+1$ frame features as input. The first M frames (context) are unmasked, while the features of the $(M+1)$-th frame (the prediction target) are entirely masked. The model is trained to regress the masked features using a Smooth L1 loss:

$$\mathcal{L}_{\text{feat}} = \text{SmoothL1}\left(G_h(h_1, \ldots, h_M), h_{M+1}\right). \tag{3}$$

At inference time, the model operates autoregressively: given M context frames, it predicts the next frame's features, which are then fed back as input for subsequent predictions. This stage ensures consistent semantic reasoning before proceeding to fine-grained synthesis in the next stage.

3.2 Semantics-Guided Video Generation

In the second stage, we use a latent video diffusion model to generate actual future frames, guided by the predicted semantic features. The model takes as input the context frames $(x_1, \ldots, x_M)$, their semantic features $(h_1, \ldots, h_M)$, and the predicted future semantic features $(h_{M+1}, \ldots, h_K)$ from $G_h(\cdot)$.

First, the context frames are encoded into compact latent features using a causal 3D VAE encoder $E_z(\cdot)$:

$$z_{1:M} = E_z(x_{1:M}). \tag{4}$$

Here, z_t has dimensions $H_z \times W_z \times C_z$, where $H_z \times W_z$ are the spatial dimensions and C_z is the number of channels. For the 3D VAE, we employ the WAN2.1 VAE [69], a causal variational autoencoder that compresses videos along both spatial and temporal dimensions. For clarity, we omit the details of temporal subsampling. To align the temporal resolution between the two stages, the semantic prediction stage processes only every $\frac{1}{r}$ frame, where r is the temporal subsampling ratio of the VAE encoder.

Next, we generate the future latent frames using a diffusion model $G_z(\cdot)$. Following standard diffusion terminology, we gradually denoise latent features for the future frames. Let $z_t^{(n)}$ denote the noisy latent of frame t at diffusion step n. The forward process gradually adds Gaussian noise to the ground-truth future latents z_t (for $t = M+1, \ldots, K$):

$$z_t^{(n)} = z_t + \sigma_n \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{5}$$

where σ_n controls the noise schedule.

The denoising model $G_z(\cdot)$ takes as input the noised future latents $z_{M+1:K}^{(n)} = (z_{M+1}^{(n)}, \ldots, z_K^{(n)})$, the clean context latents $z_{1:M} = (z_1, \ldots, z_M)$ (no noise added), all semantic features $h_{1:K} = (h_1, \ldots, h_K)$ (no noise added), and the noise step n. It predicts the clean future-frame latents $z_{M+1:K}^{(0)} = z_{M+1:K}$ for the future frames:

$$\hat{z}_{M+1:K} = G_z\big(z_{M+1:K}^{(n)};\, z_{1:M},\, h_{1:K},\, n\big). \tag{6}$$

The model is trained to minimize the denoising objective:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{n,\epsilon}\left[\lambda_n \left\|\hat{z}_{M+1:K} - z_{M+1:K}\right\|^2\right] \tag{7}$$

where λ_n is a reweighting function that balances the contribution of different noise levels. Only the future frames ($t = M+1, \ldots, K$) contribute to this loss.

Finally, the 3D VAE decoder $D_z(\cdot)$ reconstructs the pixel-space frames from the latent sequence:

$$\hat{x}_{1:K} = D_z(\hat{z}_{M+1:K}). \tag{8}$$

3.3 `Re2Pix` Architecture

The detailed architecture of the diffusion transformer during training, including early semantic alignment and nested dropout (discussed in Subsection 3.4, is shown in Figure 2.

Diffusion architecture. The diffusion model $G_z(\cdot)$ follows the video prediction design of Cosmos-Predict [1,3], which is built upon the Diffusion Transformer (DiT) framework [50]. We retain the core architectural components of Cosmos-Predict, including 3D-factorized Rotary Position Embeddings (RoPE) [60], query/key normalization before attention [13, 16, 74], and RMSNorm [85] with learnable scales in all self-attention blocks. Noise-level conditioning is implemented via LoRA-based adaptive normalization (AdaLN-LoRA) [29], which replaces the

parameter-heavy AdaLN layers of DiT, yielding a more efficient yet equally expressive design.

Compared to the original Cosmos-Predict architecture, we remove the cross-attention layers—since our model is not text-conditioned—and introduce a dedicated *semantic guidance mechanism*, described next.

Early Semantic Alignment To incorporate semantic guidance, we fuse semantic features h with the VAE latent representations z at the input level. Specifically, the VAE latents z are patchified with a spatial size of 2×2, and the semantic features h are resized to match the same spatial resolution. Both feature maps are then embedded independently and combined by channel-wise summation, ensuring joint conditioning from the outset. This design enables early semantic alignment between the predicted scene structure and the generative latent space, providing simple and efficient guidance for the diffusion process without increasing the token count.

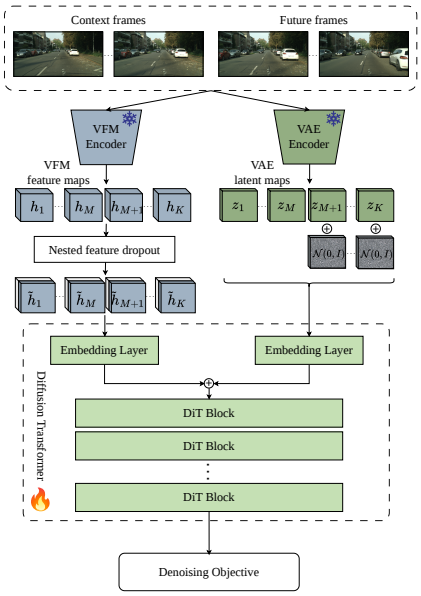

Fig. 2: `Re2Pix` architecture. The model takes as input (1) VAE latents $z_{1:K}$ with noise applied to the future frames $z_{M+1:K}$, and (2) semantic features $h_{1:K}$ from a vision foundation model (VFM), with nested dropout applied to zero-out fine-grained channels. The two modalities are embedded independently and combined via channel-wise summation at the input, implementing early fusion. The diffusion transformer is trained with a denoising objective applied to the future VAE latents. Nested dropout and mixed supervision (using for the future frames both ground-truth and predicted features with a 90%/10% mixture; not shown in the figure for simplicity) regularize the model against overfitting to idealized VFM features for the future frames.

3.4 Training Strategies for Robust Semantic Conditioning

During training, we follow a teacher-forcing approach: the semantic features h_t for future frames are extracted from the ground-truth frames using the encoder $E_h(\cdot)$. While this setup stabilizes training and accelerates convergence, it introduces a mismatch between training and inference. At test time, the semantic features are provided by G_h, which are inherently noisier than those extracted from ground truth. As a result, models trained exclusively with ground-truth features tend to overfit to these ideal representations—producing good semantic layouts but blurry pixel-level details, reflected by degraded FVD and FID scores.

To mitigate this issue, we introduce two complementary strategies that improve robustness to imperfect semantic inputs.

Nested dropout of semantic input features. The semantic features $h_t \in \mathbb{R}^{C_h \times H \times W}$ are derived from DINOv2 [49] with a ViT-B backbone, by concatenating representations from the 3rd, 6th, 9th, and 12th transformer blocks and projecting them to $C_h = 1152$ channels via PCA. These channels form a hierarchical representation, where higher-variance components capture coarse semantics and lower-variance components encode fine details.

To prevent the diffusion model from over-relying on fine-grained features, we apply nested dropout [39, 54] to all semantic feature maps $h_{1:K}$. With equal probability, we retain only the first $c \in 8, 16, 32, 64, 128, 256, 512, 1152$ channels of each h_t, replacing the remaining channels with zeros:

$$\tilde{h}_{1:K} = \text{NestedDropout}(h_{1:K}; c) = [h_{1:K}^{1:c}, \mathbf{0}^{C_h - c}], \tag{9}$$

where $\tilde{h}_{1:K}$ denotes the semantic features after nested dropout, $h_{1:K}^{1:c}$ indicates the first c feature channels retained for all frames, and $[\cdot, \cdot]$ denotes the concatenation operation. This stochastic truncation encourages the diffusion model to learn robust conditioning from the most informative semantic subspaces, rather than memorizing fine-scale correlations.

Mixed supervision with ground-truth and predicted semantic features. To further reduce the train–test gap, we train the diffusion model with a mixture of ground-truth and predicted semantic features. For each batch, we randomly sample 10% of examples where the semantic features are taken from the generator G_h (predicted features), and 90% where they come from E_h (ground-truth features). This mixture regularization exposes the diffusion model to both ideal and imperfect semantic inputs during training. Empirically, the 90/10 ratio provides the best trade-off—reducing blur and improving visual quality without compromising semantic consistency.

4 Experiments

Our experiments assess how well the proposed hierarchical `Re2Pix` framework predicts future video frames that are both semantically faithful and photorealistic. We evaluate on multiple datasets and compare against strong baselines, analyze training dynamics (in Subsection 4.2), and conduct detailed ablations (in Subsection 4.3). Across metrics, our method yields consistent gains in temporal semantic consistency, generation quality, and training efficiency, demonstrating the effectiveness of coarse-to-fine hierarchical visual modeling.

4.1 Experimental Setup

Datasets. We experiment on four driving datasets: Cityscapes [12], nuScenes [10], CoVLA [4], and KITTI [20]. Our primary setup trains and evaluates on Cityscapes.

To assess generalization at larger scale, we additionally train on a combination of Cityscapes, nuScenes, and CoVLA, evaluating in-domain on Cityscapes and nuScenes, and zero-shot on KITTI. Full dataset details are provided in Appendix Section 8.

Implementation Details. We use DINOv2-Reg ViT-B/14 [49] as the VFM encoder for semantic feature extraction. Our diffusion transformer follows the EDM formulation [33] used in Cosmos-Predict [1], with 14 layers, 16 attention heads, and a 2048-dimensional embedding, totaling roughly 800M parameters.

We train on sequences of $K = 25$ frames, where $M = 13$ context frames at 432×768 resolution are encoded using the WAN2.1 VAE [69] with $8 \times 8 \times 4$ temporal-spatial downsampling, yielding 7 latent frames of size 54×96. Models are trained from scratch using Adam [36] ($\beta_1 = 0.9$, $\beta_2 = 0.99$) with a learning rate of $0.6 \times 2^{-10.5}$ and linear warmup and decay. Training runs for 40k iterations on 8 NVIDIA H200 GPUs for single dataset experiments and 120k iterations for multiple dataset experiments) with effective batch size 8, requiring approximately 7 and 28 hours respectively. Full implementation details regarding diffusion formulation, feature prediction model and semantic decoders are provided in Appendix Section 9.

Evaluation Protocol and Metrics. For all experiments, we use frames 3–15 as context and predict frames 16–27. We evaluate two complementary aspects:

- **Temporal semantic consistency:** How well the generated future frames preserve the evolving scene semantics.
- **Generation quality:** How realistic and temporally coherent the synthesized frames appear.

Semantic consistency metrics. We evaluate semantic segmentation (mIoU) and depth estimation on the generated frame 19. Segmentation is computed using a DINOv2-Reg ViT-B encoder with DPT heads [53]. We report mIoU over all classes (A) and over moving-object classes (M), which are more challenging and critical for autonomous driving scenarios. Depth evaluation uses Absolute Relative Error (AbsRel) and threshold accuracy (δ_1), where lower AbsRel and higher δ_1 indicate better performance. Following standard practice, frame 19 in each sequence provides dense annotations for 19 semantic classes. Since Cityscapes and nuScenes not include dense depth labels, we generate pseudo-depth using Depth Anything V2 [78].

Generation quality metrics. We compute FID [25] and FVD [64] over all predicted frames to capture both spatial realism and temporal coherence. Lower scores indicate better generative quality.

4.2 Video Prediction Results

Comparison with Baselines We compare our hierarchical approach with three baselines: (i) a standard diffusion video model trained with a denoising objective,

Table 1: Comparison with baselines. All numbers report absolute performance; values in parentheses indicate improvement over the Baseline. Blue indicates improvement, red indicates degradation. Our hierarchical approach provides consistent gains across both semantic consistency and generation quality. Model parameters: Baseline (782M), REPA variants (792M), Baseline-Large (1.5B), `Re2Pix` (1.1B).

Method	Semantic Consistency				Generation	
	mIoU(A)↑	IoU(M)↑	δ_1↑	AbsR↓	FID↓	FVD↓
`Re2Pix` (Stage-1)	69.76	69.66	88.03	.1262	-	-
Baseline	60.55	57.64	85.15	.1460	12.86	60.70
w/ REPA [41]	61.45 (+0.90)	59.63 (+1.99)	85.35 (+0.20)	.1465 (+0.0005)	12.34 (-0.52)	55.15 (-5.55)
w/ VideoREPA [86]	60.98 (+0.43)	58.61 (+0.97)	85.22 (+0.07)	.1466 (+0.0006)	12.95 (+0.09)	59.03 (-1.67)
Baseline-Large	61.69 (+1.14)	59.65 (+2.01)	85.43 (+0.28)	.1453 (-0.0070)	11.99 (-0.87)	56.81 (-3.89)
`Re2Pix`	**63.53 (+2.98)**	**62.29 (+4.65)**	**85.72 (+0.57)**	**.1413 (-0.0047)**	**9.90 (-2.96)**	**52.66 (-8.04)**

(ii) REPA [84], and (iii) VideoREPA [86]. For reference, we additionally report the semantic consistency results of the VFM feature forecasting model used at stage 1 as `Re2Pix` (Stage 1).

As shown in Table 1, our `Re2Pix` method achieves substantial improvements across both semantic consistency and generation quality. Segmentation and depth metrics show that `Re2Pix` produces future frames with significantly improved temporal semantic fidelity. At the same time, `Re2Pix` achieves the best FID and FVD scores, demonstrating its ability to synthesize photorealistic and temporally coherent predictions. These results confirm that jointly modeling semantic structure and pixel-space generation through a hierarchical design provides strong performance gains over existing methods. To account for the stochastic nature of our diffusion model, we additionally report mean and standard deviation over 3 independent sampling runs in *Appendix Subsection 6.2*.

Does `Re2Pix` simply benefit from more parameters? To answer this question, we train a stronger baseline with additional diffusion transformer layers, resulting in a parameter count exceeding that of `Re2Pix`. As shown in Table 1, increasing the parameter count improves baseline performance but remains consistently inferior to `Re2Pix` on both semantic and generation metrics. This indicates that the improvements arise from the hierarchical strategy rather than model size.

`Re2Pix` accelerates diffusion training. Recent representation-alignment methods for images [84] have shown that semantically aligned guidance can improve convergence. We test whether our hierarchical approach offers similar benefits in video generation. To study convergence dynamics, we train both the baseline and `Re2Pix` for 140k iterations using a constant learning rate after warmup, enabling direct comparison across checkpoints. As shown in Figure 3, `Re2Pix` outperforms the baseline throughout training for all metrics:

- For FID, `Re2Pix` reaches 15 in **20k iterations**, whereas the baseline requires **140k**—a 7× **speed-up**. FVD exhibits a similar 7× **acceleration**.
- Semantic metrics also converge faster: segmentation mIoU achieves a 14× **speed-up**.

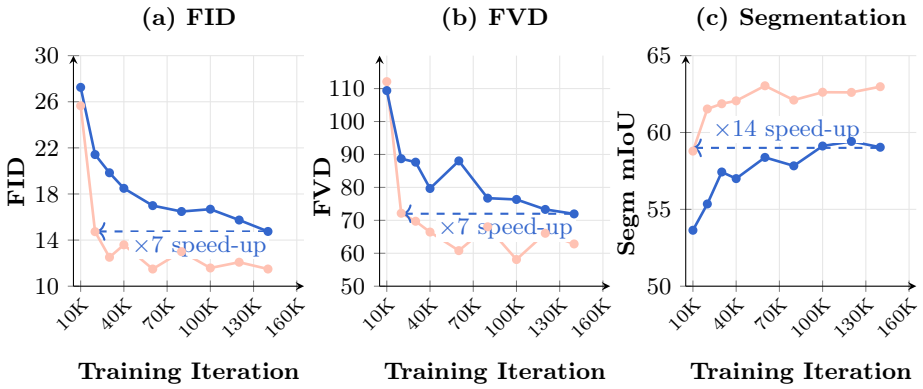

Fig. 3: Accelerated Training Convergence. Training curves for (a) FID, (b) FVD, and (c) Segmentation mIoU comparing our hierarchical approach (orange) with the baseline (blue). Our method achieves ×7 speed-up for generation metrics and ×14 speed-up for segmentation.

Qualitative comparisons across diverse driving scenes are provided in Appendix Section 7.

Computational cost. Table 2 summarizes the inference and training cost of all methods. Re2Pix introduces negligible inference overhead over the Baseline (3.2s/28.5GB vs. 3.1s/27GB), while remaining considerably more efficient than Baseline-Large (4.8s/28.8GB). Although Re2Pix requires slightly more training time per iteration (7h vs. 5.3h for 40k iterations), the 7-14× convergence speedup demonstrated in Figure 3 makes the additional per-iteration cost negligible in practice.

Table 2: Computational cost. Inference time and peak memory per scene, and training time for 40k iterations.

Method	Inf. Time↓	Inf. Mem.↓	Train. Time↓	Train. Mem.↓
Baseline	3.1s	27.0GB	5.3h	36GB
Baseline-Large	4.8s	28.8GB	8.0h	41GB
Re2Pix	3.2s	28.5GB	7.0h	40GB

Scaling and Cross-Dataset Generalization. To evaluate whether the benefits of our hierarchical framework scale with data diversity, we train `Re2Pix` on a larger combined dataset (Cityscapes, nuScenes, and CoVLA) and assess performance across both in-domain (Cityscapes and nuScenes) and zero-shot (KITTI) settings. As shown in Table 3, our method consistently outperforms the baselines (Baseline and Baseline Large) across all benchmarks: on Cityscapes and nuScenes, it improves both semantic consistency and generation quality, while on KITTI, it demonstrates superior zero-shot generalization. We further compare our approach against state-of-the-art large-scale systems, specifically Vista [18] and Cosmos-Predict 2 [48]. Although these models are pretrained with several orders of magnitude more data and compute—and were subsequently fine-tuned in our specific setting—`Re2Pix` surpasses Vista and performs competitively with Cosmos-Predict 2. This result is particularly significant as it demonstrates that our hierarchical design achieves comparable—and in some cases superior—results with a fraction of the training overhead.

Table 3: *Results for* ***Re2Pix*** *trained on extended data (Cityscapes + nuScenes + CoVLA). We report in-domain results on Cityscapes and nuScenes, zero-shot generalization on KITTI, and comparisons against* ***Re2Pix*** *(Stage 1), Baselines and large-scale internet-pretrained systems (Vista, Cosmos-Predict-2), both fine-tuned in our setting. For reference, we include* ***Re2Pix*** *(Stage 1), which reports semantic consistency metrics only, as it operates purely in feature space without the ability of generating pixels.*

Method	Cityscapes						nuScenes				KITTI	
	Semantic Consistency				Generation		Sem. Cons.		Generation		Generation	
	mIoU(A)↑	IoU(M)↑	δ_1↑	AbsR↓	FID↓	FVD↓	δ_1↑	AbsR↓	FID↓	FVD↓	FID↓	FVD↓
VFM forecasting only												
Re2Pix (Stage 1)	70.46	70.26	88.03	.1208	-	-	83.82	.1869	-	-	-	-
Trained from scratch												
Baseline	62.77	60.86	85.62	.1433	9.64	52.12	79.92	.2438	21.08	136.71	20.02	**61.02**
Baseline (Large)	63.32	61.73	85.63	.1423	9.89	51.77	80.17	.2412	20.73	134.58	20.94	61.28
Re2Pix (ours)	**64.63**	**63.52**	**86.01**	**.1400**	**9.29**	**49.03**	**81.09**	**.2306**	**18.96**	**134.03**	**15.94**	61.73
Large-scale pretrained												
Vista (Finetuned)	63.88	62.35	85.91	.1376	12.17	84.14	80.46	.2304	21.06	174.86	7.95	97.96
Cosmos-Predict-2 (Finetuned)	65.69	64.85	86.19	.1371	7.74	46.22	81.28	.2243	15.90	133.19	8.62	41.64

We additionally evaluate Re2Pix on a non-driving dataset of non-rigid human-object interactions in Appendix Subsections 6.1. These results demonstrate that incorporating hierarchical semantic guidance not only improves final performance but also substantially accelerates training.

4.3 Ablation Study

Effect of Nested Dropout. We evaluate the impact of nested dropout by comparing two settings (Table 4): using a fixed set of 1152 feature channels, and applying nested dropout during training (with the full 1152 at test time). Nested dropout yields consistent improvements across all metrics. The largest gains appear in FID and FVD, indicating substantially better generation quality. By stochastically truncating fine-grained semantic channels during training, nested dropout encourages the diffusion model to rely on robust, coarse-to-fine semantic structure rather than overfitting to perfect ground-truth feature values. This reduces the train–test mismatch between ground-truth features (seen during training) and predicted features (used at inference), enabling sharper and more realistic synthesis without degrading semantic fidelity. Building on nested dropout, we further investigate a CFG-inspired representation guidance scheme that contrasts predictions at different levels of semantic granularity. Ablations on this technique property are provided Appendix Subsection 6.3.

Sensitivity to VFM features. To assess whether Re2Pix depends on a specific vision foundation model, we replace DINOv2 with SigLIP-2 [62] as the feature extractor in Stage 1 and retrain on Cityscapes (Table 5). Both variants consistently outperform the Baseline across semantic consistency and generation metrics, demonstrating that the hierarchical design is robust to the choice of VFM, with DINOv2 yielding slightly stronger results consistent.

Mixed supervision with ground-truth and predicted features. We compare three training configurations (Table 6): (i) conditioning on ground-truth VFM fea-

Table 4: Impact of Nested Dropout. Training with nested dropout (bottom) substantially improves generation quality compared to training with fixed 1152 components (top), while also improving the semantic consistency metrics.

Components	Sem. Consistency				Generation	
	mIoU(A)↑	IoU(M)↑	δ_1↑	AbsR↓	FID↓	FVD↓
Fixed (1152)	62.38	60.67	85.45	.1467	12.63	58.91
Nested	**63.53**	**62.29**	**85.72**	**.1413**	**9.90**	**52.66**

Table 5: Sensitivity to VFM features. We replace DINOv2 with SigLIP-2 as the feature extractor in Stage 1 and report results on Cityscapes. Both variants consistently outperform the Baseline, demonstrating that the hierarchical design is robust to the choice of VFM

Method	Semantic Consistency				Generation	
	mIoU(A)↑	IoU(M)↑	δ_1↑	AbsR↓	FID↓	FVD↓
Baseline	60.55	57.64	85.15	.1460	12.86	60.70
`Re2Pix` w/ SigLIP2	63.02	61.51	85.73	.1419	10.26	52.69
`Re2Pix` w/ DinoV2	**63.53**	**62.29**	**85.72**	**.1413**	**9.90**	**52.66**

Table 6: Impact of mixed supervision with GT and predicted features. Training with a mixture of ground-truth (90%) and predicted features (10%) combines the semantic consistency benefits of training with ground-truth features with the generation quality benefits of training with predicted features.

Features	Sem. Consistency				Generation	
	mIoU(A)↑	IoU(M)↑	δ_1↑	AbsR↓	FID↓	FVD↓
Baseline	60.55	57.64	85.15	.1460	12.86	60.70
Ground Truth	64.42	63.15	86.06	.1407	10.43	80.85
Predicted	62.77	61.38	85.58	.1420	10.21	55.81
Mixed (85/15)	62.49	60.00	85.52	.1440	10.54	54.89
Mixed (90/10)	**63.53**	62.29	85.72	.1413	**9.90**	**52.66**
Mixed (95/05)	63.49	**62.33**	**85.75**	**.1319**	10.27	53.69

Table 7: Impact of fusion strategy. Comparison of early fusion, AdaLN conditioning, and cross-attention on Cityscapes. Early fusion achieves the best performance across both semantic consistency and generation metrics while introducing no additional parameters or architectural complexity.

Features	Sem. Consistency				Generation	
	mIoU(A)↑	IoU(M)↑	δ_1↑	AbsR↓	FID↓	FVD↓
Baseline	60.55	57.64	85.15	.1460	12.86	60.70
Early-Fusion	**63.53**	**62.29**	**85.72**	**.1413**	**9.90**	**52.66**
AdaLN	63.10	61.75	85.71	.1425	10.75	55.75
Cross-Attn	61.34	59.03	85.30	.1462	11.83	56.37

tures, (ii) conditioning on predicted features, and (iii) mixed strategies with varying ground-truth/predicted ratios (85/15, 90/10, 95/05).

Training only on ground-truth features yields the strongest semantic consistency, but severely degrades perceptual quality due to a train–test mismatch: the model overfits to precise ground-truth feature values for generating fine-grained image details and struggles when exposed to noisier predicted features at inference, producing blurry frames. Training only on predicted features reverses this trend—generation metrics improve, but perception metrics degrade.

Among the mixed strategies, the 90/10 ratio achieves the best overall trade-off, with more extreme ratios degrading either semantic consistency (85/15) or generation quality (95/05), confirming that stochastically mixing the two sources at the right ratio enables robust semantic conditioning while preserving high-fidelity synthesis.

Fusion strategy. We adopt early fusion as a principled yet minimal design choice, combining VAE latents and VFM features via channel-wise summation at the input level. To validate this choice, we ablate two alternative conditioning strategies on Cityscapes (Table 7): AdaLN, which reuses the existing time-conditioning

layers also for semantic features, and cross-attention, which introduces additional layers at a cost of 265M extra parameters. Early fusion outperforms both alternatives across semantic consistency and generation metrics, confirming that simple input-level fusion provides strong and efficient semantic guidance without additional architectural complexity.

Number of semantic components at inference. Because nested dropout exposes the model to varying semantic feature dimensionalities during training, we can adjust the number of PCA components at inference time. As shown in Table 8, performance remains strong even when reducing to 128 components, indicating that coarse semantic structure captured by top principal components is highly informative. Performance degrades gracefully at very low dimensions (e.g., 8 or 16 components), while the full 1152 components yield the best scores w.r.t. the semantic consistency metrics.

Table 8: Number of semantic components at inference. Performance across va numbers of PCA components at inference. Using 256 components achieves comparable results to the full 1152 components.

Components	Sem. Consistency				Generation	
	mIoU(A)↑	IoU(M)↑	δ_1↑	AbsR↓	FID↓	FVD↓
8	62.07	60.33	85.52	.1441	10.44	54.20
16	62.56	60.88	85.51	.1436	10.22	55.05
32	63.09	61.81	85.59	.1431	9.95	54.16
64	63.19	61.75	85.67	.1420	9.94	53.34
128	63.27	61.72	85.56	.1438	**9.80**	52.82
256	63.44	62.10	**85.80**	.1417	9.88	52.46
512	63.46	62.22	85.67	.1430	9.95	**52.00**
1152	**63.53**	**62.29**	85.72	**.1413**	9.90	52.66

5 Conclusion

We introduced `Re2Pix`, a hierarchical semantic-to-pixel framework for video prediction that integrates VFM representation forecasting with a diffusion-based video generator. By first predicting future semantic representations and then leveraging them to guide pixel-space synthesis, `Re2Pix` produces videos that are semantically consistent, temporally coherent, and photorealistic. Extensive experiments on multiple datasets demonstrate substantial improvements over strong baselines, including REPA, across temporal semantic consistency, perceptual quality, and training efficiency, with significant acceleration in convergence. Ablation studies further highlight the importance of nested dropout and mixed supervision for robust semantic conditioning and high-fidelity generation. Overall, our results show that explicitly modeling hierarchical semantic structure is an effective and scalable strategy for generating realistic future video frames, paving the way for more reliable video prediction in complex dynamic scenes.

Acknowledgements This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. Hardware resources were granted with the support of GRNET. Also, this work was performed using EuroHPC resources (Project ID e-dev-2026d01-061) and HPC resources from GENCI-IDRIS (Grants AD011016639 and AS011017163).

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025) 4, 7, 10
2. Alhaija, H.A., Alvarez, J., Bala, M., Cai, T., Cao, T., Cha, L., Chen, J., Chen, M., Ferroni, F., Fidler, S., et al.: Cosmos-transfer1: Conditional world generation with adaptive multimodal control. arXiv preprint arXiv:2503.14492 (2025) 4
3. Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., et al.: World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062 (2025) 4, 7
4. Arai, H., Miwa, K., Sasaki, K., Watanabe, K., Yamaguchi, Y., Aoki, S., Yamamoto, I.: Covla: Comprehensive vision-language-action dataset for autonomous driving. In: WACV. pp. 1933–1943. IEEE (2025) 9
5. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025) 4
6. Baldassarre, F., Szafraniec, M., Terver, B., Khalidov, V., Massa, F., LeCun, Y., Labatut, P., Seitzer, M., Bojanowski, P.: Back to the features: Dino as a foundation for video world models. arXiv preprint arXiv:2507.19468 (2025) 4
7. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. Transactions on Machine Learning Research (2024) 4
8. Bartoccioni, F., Ramzi, E., Besnier, V., Venkataramanan, S., Vu, T.H., Xu, Y., Chambon, L., Gidaris, S., Odabas, S., Hurych, D., et al.: Vavim and vavam: Autonomous driving through video generative modeling. arXiv preprint arXiv:2502.15672 (2025) 4
9. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 4
10. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020) 9
11. Chen, H., Han, Y., Chen, F., Li, X., Wang, Y., Wang, J., Wang, Z., Liu, Z., Zou, D., Raj, B.: Masked autoencoders are effective tokenizers for diffusion models. In: ICML (2025) 5
12. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (June 2016) 9
13. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. In: International conference on machine learning. PMLR (2023) 7
14. Denton, R., Fergus, R.: Stochastic video generation with a learned prior. In: ICML (2018) 4
15. Dosovitskiy, A., Koltun, V.: Learning to act by predicting the future. In: ICLR (2017) 1

16. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024) 7
17. Feng, T., Wang, W., Yang, Y.: A survey of world models for autonomous driving. arXiv preprint arXiv:2501.11260 (2025) 1
18. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. In: NeurIPS (2024) 1, 4, 12
19. Gao, Z., Tan, C., Wu, L., Li, S.Z.: Simvp: Simpler yet better video prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3170–3180 (June 2022) 4
20. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research (2013) 9
21. Guan, Y., Liao, H., Li, Z., Hu, J., Yuan, R., Zhang, G., Xu, C.: World models for autonomous driving: An initial survey. IEEE Transactions on Intelligent Vehicles (2024) 1
22. Gupta, A., Tian, S., Zhang, Y., Wu, J., Martín-Martín, R., Fei-Fei, L.: Maskvit: Masked visual pre-training for video prediction. In: ICLR (2023) 4
23. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 4
24. He, Y., et al.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint (2022) 4
25. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) NeurIPS. vol. 30. Curran Associates, Inc. (2017) 10
26. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022) 1
27. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: NeurIPS (2022), arXiv:2204.03458 4
28. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023) 5
29. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR (2022) 7
30. Hu, J.F., Sun, J., Lin, Z., Lai, J.H., Zeng, W., Zheng, W.S.: Apanet: Auto-path aggregation for future instance segmentation prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 4
31. Hwang, S., Jang, H., Kim, K., Park, M., Choo, J.: Cross-frame representation alignment for fine-tuning video diffusion models. arXiv preprint arXiv:2506.09229 (2025) 5
32. Jin, X., Li, X., Xiao, H., Shen, X., Lin, Z., Yang, J., Chen, Y., Dong, J., Liu, L., Jie, Z., et al.: Video scene parsing with predictive feature learning. In: ICCV (2017) 4
33. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. NeurIPS **35**, 26565–26577 (2022) 10
34. Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Dino-foresight: Looking into the future with dino. NeurIPS **38**, 163779–163811 (2025) 4, 6
35. Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Advancing semantic future prediction through multimodal visual sequence transformers. CVPR (2025) 4

36. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 10
37. Kouzelis, T., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Eq-vae: Equivariance regularized latent space for improved generative image modeling. ICML (2025) 4
38. Kouzelis, T., Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Boosting generative image modeling via joint image-feature synthesis. In: NeurIPS (2025) 5
39. Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., et al.: Matryoshka representation learning. NeurIPS (2022) 9
40. Lee, A.X., Zhang, R., Namee, B., et al.: Stochastic adversarial video prediction. In: arXiv preprint (2018) 4
41. Leng, X., Singh, J., Hou, Y., Xing, Z., Xie, S., Zheng, L.: Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. arXiv preprint arXiv:2504.10483 (2025) 5, 11
42. Li, T., Katabi, D., He, K.: Return of unconditional generation: A self-supervised representation generation method. NeurIPS (2024) 5
43. Li, X., Qiu, K., Chen, H., Kuen, J., Gu, J., Raj, B., Lin, Z.: Imagefolder: Autoregressive image generation with folded tokens. arXiv preprint arXiv:2410.01756 (2024) 5
44. Lin, Z., Sun, J., Hu, J.F., Yu, Q., Lai, J.H., Zheng, W.S.: Predictive feature learning for future segmentation prediction. In: CVPR (2021) 4
45. Mallya, A., Wang, T.C., Sapra, K., Liu, M.Y.: World-consistent video-to-video synthesis. arXiv preprint arXiv:2007.08509 (2020) 4
46. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv preprint (2016) 4
47. Nabavi, S.S., Rochan, M., Wang, Y.: Future semantic segmentation with convolutional lstm. In: BMVC (2018) 4
48. NVIDIA: Cosmos-predict2. https://github.com/nvidia-cosmos/cosmos-predict2 (2025), accessed: 2026-03-05 4, 12
49. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024) 3, 4, 6, 9, 10
50. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023) 1, 3, 7
51. Pernias, P., Rampas, D., Richter, M.L., Pal, C.J., Aubreville, M.: Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In: ICLR (2024) 5
52. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024) 4
53. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: CVPR (2021) 10
54. Rippel, O., Gelbart, M., Adams, R.: Learning ordered representations with nested dropout. In: ICML (2014) 9
55. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 4

56. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025) 5
57. Saric, J., Orsic, M., Antunovic, T., Vrazic, S., Segvic, S.: Warp to the future: Joint forecasting of features and feature motion. In: CVPR (2020) 4
58. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025) 4
59. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: CVPR (2021) 4
60. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing (2024) 7
61. Sun, J., Xie, J., Hu, J.F., Lin, Z., Lai, J., Zeng, W., Zheng, W.S.: Predicting future instance segmentation with contextual pyramid convlstms. In: Proceedings of the 27th ACM International Conference on Multimedia (2019) 4
62. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 4, 13
63. Tu, S., Zhou, X., Liang, D., Jiang, X., Zhang, Y., Li, X., Bai, X.: The role of world models in shaping autonomous driving: A comprehensive survey. arXiv preprint arXiv:2502.10498 (2025) 1
64. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018) 10
65. Venkataramanan, S., Pariza, V., Salehi, M., Knobel, L., Gidaris, S., Ramzi, E., Bursuc, A., Asano, Y.M.: Franca: Nested matryoshka clustering for scalable visual representation learning. arXiv preprint arXiv:2507.14137 (2025) 4
66. Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction. In: ICLR (2017) 4
67. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: CVPR (2016) 4
68. Walker, J.C., Vélez, P., Cabrera, L.P., Zhou, G., Kabra, R., Doersch, C., Ovsjanikov, M., Carreira, J., Ginosar, S.: Generalist forecasting with frozen video models via latent diffusion. arXiv preprint arXiv:2507.13942 (2025) 4
69. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 3, 4, 7, 10
70. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. NeurIPS (2018) 4
71. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024) 4
72. Wang, Y., Jiang, L., Yang, M.H., Li, L.J., Long, M., Fei-Fei, L.: Eidetic 3d LSTM: A model for video prediction and beyond. In: ICLR (2019) 4
73. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: CVPR. pp. 14749–14759 (June 2024) 1
74. Wortsman, M., Liu, P.J., Xiao, L., Everett, K., Alemi, A., Adlam, B., Co-Reyes, J.D., Gur, I., Kumar, A., Novak, R., et al.: Small-scale proxies for large-scale transformer training instabilities. arXiv preprint arXiv:2309.14322 (2023) 7

75. Wu, G., Zhang, S., Shi, R., Gao, S., Chen, Z., Wang, L., Chen, Z., Gao, H., Tang, Y., Yang, J., et al.: Representation entanglement for generation: Training diffusion transformers is much easier than you think. NeurIPS (2025) 5
76. Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., Bian, J.: Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. arXiv preprint arXiv:2507.07982 (2025) 5
77. Yan, W., Zhang, Y., et al.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint (2021) 4
78. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: NeurIPS (2024) 10
79. Yang, S., et al.: Video diffusion models with local-global context guidance. arXiv preprint (2023) 4
80. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024) 4
81. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: CVPR (2025) 5
82. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: CVPR (2023) 4
83. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., Gong, B., Yang, M.H., Essa, I., Ross, D.A., Jiang, L.: Language model beats diffusion - tokenizer is key to visual generation. In: ICLR (2024) 4
84. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: ICLR (2025) 2, 3, 5, 11
85. Zhang, B., Sennrich, R.: Root mean square layer normalization. NeurIPS (2019) 7
86. Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., Cheng, Y.: VideoREPA: Learning physics for video generation through relational alignment with foundation models. In: NeurIPS (2025) 2, 3, 5, 11
87. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025) 5
88. Zhou, G., Pan, H., LeCun, Y., Pinto, L.: DINO-WM: World models on pre-trained visual features enable zero-shot planning. In: ICML (2025) 4