SyncVP: Joint Diffusion for Synchronous Multi-Modal Video Prediction

Enrico Pallotta, Sina Mokhtarzadeh Azar, Shuai Li, Olga Zatsarynna, Juergen Gall
University of Bonn, Lamarr Institute for Machine Learning and Artificial Intelligence
{pallotta,mokhtarzadeh,lishuai,zatsarynna,gall}@iai.uni-bonn.de

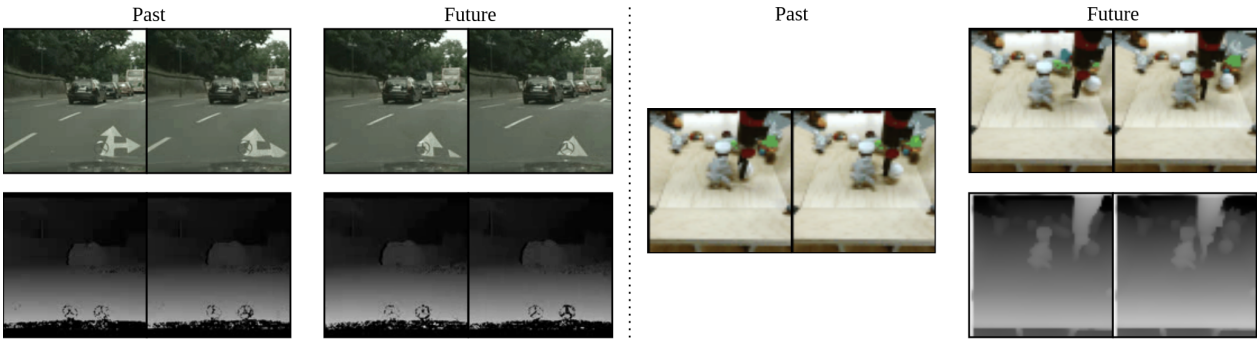


Figure 1. SyncVP is a diffusion model for synchronized multi-modal video prediction. It generates multi-modal future frames like RGB and depth for a given observation that can consist of both modalities (left) or only one modality (right).

Abstract

Predicting future video frames is essential for decision-making systems, yet RGB frames alone often lack the information needed to fully capture the underlying complexities of the real world. To address this limitation, we propose a multi-modal framework for Synchronous Video Prediction (SyncVP) that incorporates complementary data modalities, enhancing the richness and accuracy of future predictions. SyncVP builds on pre-trained modality-specific diffusion models and introduces an efficient spatio-temporal cross-attention module to enable effective information sharing across modalities. We evaluate SyncVP on standard benchmark datasets, such as Cityscapes and BAIR, using depth as an additional modality. We furthermore demonstrate its generalization to other modalities on SYNTHIA with semantic information and ERA5-Land with climate data. Notably, SyncVP achieves state-of-the-art performance, even in scenarios where only one modality is present, demonstrating its robustness and potential for a wide range of applications.

1. Introduction

Video prediction, the task of forecasting future video frames based on past video frames, has gained significant attention in recent years [3, 39, 42, 44] due to its broad range of applications. Autonomous driving, weather forecasting, healthcare and human-machine interaction are just a few examples of scenarios in which the ability to anticipate future events is critical. In these contexts, accurate video prediction enables systems to react and adapt in real-time, enhancing both safety and efficiency or providing valuable information for decision-making processes. While traditional video prediction focuses on generating future frames of RGB videos, many real-world applications involve multiple sensory inputs and require a deep understanding of the underlying real world dynamics. This leads to a natural extension of the task to the multi-modal domain, where additional data such as depth or semantic information are used alongside RGB frames. Several works have thus investigated how to exploit other modalities for conditioning the generation of images or videos, like text [5, 14, 28, 36, 48, 58], depth [13, 21, 24, 49, 57], pose [29, 57], flow [32], sketches [43] and audio [35, 50]. Although significant progresses have been made in guiding video generation through the use of auxiliary modalities, few studies have focused on developing models capable of leveraging multiple modalities while simultaneously generating each of them.

In this paper, we thus propose an approach for multi-modal video prediction as shown in Fig. 1. Given a set of observed frames, the model predicts future frames of two modalities like RGB and depth in a consistent manner. While our approach is based on a latent diffusion model [56], we demonstrate that a naive concatenation of

the modalities performs poorly since each modality has its own characteristics. Instead, we propose a pair of synchronized diffusion models that share relevant information through an efficient cross-attention mechanism, which works separately on the spatial and temporal dimensions. Another key aspect is the noise sharing between the modalities, which not only ensures synchronous predictions across the modalities but also has a very positive impact on the loss convergence. Finally, we propose a novel cross-modality guidance approach for training, which enables the model to predict multi-modal frames even if only one modality is observed as shown in Fig. 1. We summarize the contributions of our *Synchronous multi-modal Video Prediction* framework (SyncVP) in the following key points:

- A generalized and scalable multi-modal framework for video prediction that can exploit pre-trained modality specific diffusion models with little finetuning.
- We introduce a lightweight and efficient module for cross-modality information exchange through a split spatio-temporal cross-attention mechanism that computes only a single shared attention matrix.
- We propose to use a shared forward diffusion process by applying the same noise to each modality. This method leads to significant improvements in loss convergence and conditional generation performance compared to using independent noises.
- We propose *cross-modality guidance*, a joint modality training technique that enables simultaneous multi-modal video prediction also with partial conditioning. By randomly masking one or both modality inputs during training, the model learns to predict future frames even when one modality is missing.

2. Related Work

We discuss related works for video prediction and multi-modal generation, as this work bridges these two tasks and establishes a multi-modal framework for synchronous video prediction.

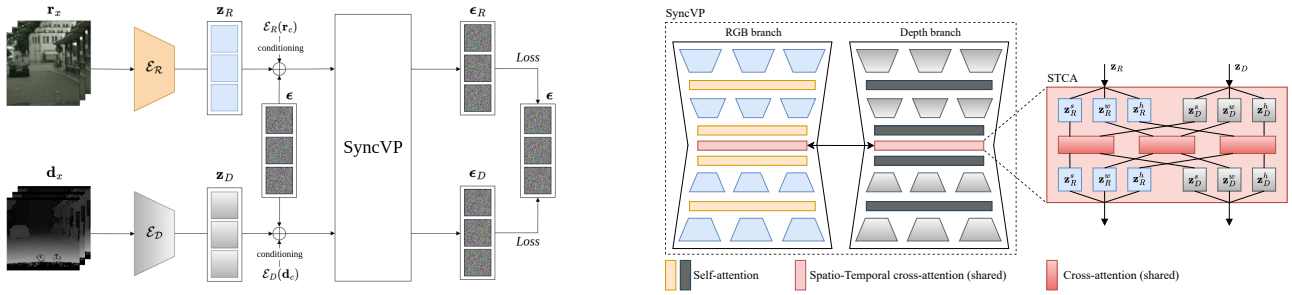
2.1. Video prediction

Early approaches for video frames prediction primarily focused on recurrent networks like ConvLSTM and RNNs [6, 40, 44]. Due to the inherent uncertain nature of future frames in a video, several works proposed stochastic methods based on variational models (VAEs, VRNNs) [1–4, 7, 11, 41, 47]. GANs have also proven effective in video prediction [9, 22, 27, 39]. Other approaches tackled this task as a neural process [54] mapping input spatio-temporal coordinates to target pixel values, using tailored video Transformers [53] or, inspired by the human vision system [19], defining a model for the frequency domain through the use of a multi-level wavelet transform. In the last three years, Diffusion models have been dominating the field of image generation [16, 33], and they have shown outstanding performances on videos [5, 17]. Based on the diffusion paradigm, several works proposed models for video prediction [18, 26, 42, 55, 59], each of them proposing different backbone models, past frames conditioning techniques or ways to model the spatio-temporal dependencies. Although these studies have progressively improved the quality of generated videos, none has yet focused on the integration and exploitation of multi-modal information.

2.2. Multi-modal generation

Multi-modal generation is the task of generating semantically aligned outputs across different data modalities. In the image domain, we find several works that aim to learn the joint distribution of multiple modalities. LDM3D [38] trains a diffusion model for generating an RGB+D image from text by concatenating the RGB and depth images along the channel dimension before feeding the resulting vector to the latent encoder. MT-Diffusion [8] instead defines a common diffusion space by aggregating the latent representations of each modality, which are separately encoded. DiffX [45] proposes a multi-path VAE to encode and decode all the modalities in a single shared latent space before training a diffusion model for layout-guidance. HyperHuman [25] uses a joint diffusion UNet with expert branches of RGB, depth and surface normal for image generation. The generated depth and normal images are then used in a second refining step as conditioning for generating an RGB image with higher resolution. Finally, Xing et al. [50] propose an inference time optimization technique that uses a pre-trained ImageBind model to align the latent vectors generated by two separate video and audio diffusion models.

Focusing on video models, MM-Diffusion [35] employs two coupled denoising networks for joint audio-video generation. The whole system is trained in a single step with a cross-attention mechanism that synchronizes the two modalities. Similarly, IDOL [57] also uses two denoising UNets coupled with cross-attention for pose-guided human image animation. They first estimate the depth map for the given image and then generate an animation of the two modalities. They furthermore introduce loss terms for motion and cross-attention maps to enhance the consistency between generated RGB and depth frames. Although CVD [20] does not handle multiple modalities, they propose a video generative diffusion model for multiple views. A cross-view attention module is trained on top of a Stable Diffusion model to condition on different camera poses. To the best of our knowledge, this is the first work addressing multi-modal generation in a pure video prediction manner, defining a multi-modal video-to-video framework that exploits complementary information from all the modalities to improve the generation quality of the predicted video.



(a) Joint diffusion training pipeline

(b) SyncVP architecture and cross-attention module

Figure 2. Our multi-modal latent diffusion framework is trained by exploiting pre-trained weights (θ_R, θ_D) of each modality and an efficient spatio-temporal cross-attention mechanism. The same noise $\boldsymbol{\epsilon}$ is used during the forward diffusion process of each modality.

3. Synchronous Multi-Modal Video Prediction

Our goal is to forecast multi-modal video frames from a short sequence of observed frames where we focus on RGB and depth as modalities. As illustrated in Fig. 1, we propose an approach with a novel cross-modality guidance, i.e., the multi-modal model can be conditioned on both or only one modality. Before we describe the approach in detail, we first introduce a formal definition of joint multi-modal video prediction. Given a dataset of paired RGB and depth videos $D = \{(\mathbf{r}_i, \mathbf{d}_i)\}_{i=1}^{N}$, our goal is to predict P future RGB $\mathbf{r}_x = (r_{t_1}, r_{t_2}, ..., r_{t_P})$ and depth frames $\mathbf{d}_x = (d_{t_1}, d_{t_2}, ..., d_{t_P})$ given C past RGB $\mathbf{r}_c = (r_{t_1}, r_{t_2}, ..., r_{t_C})$ and depth frames $\mathbf{d}_c = (d_{t_1}, d_{t_2}, ..., d_{t_C})$. We are therefore interested in learning the joint conditional distribution $p(\mathbf{r}_x, \mathbf{d}_x \mid \mathbf{r}_c, \mathbf{d}_c)$. While we describe the proposed Synchronous multi-modal Video Prediction (SyncVP) framework, which is based on a novel spatio-temporal cross-attention, in Sec. 3.1, Sec. 3.2 describes the training using cross-modality guidance.

3.1. Spatio-temporal cross-attention

Our proposed Synchronous multi-modal Video Prediction (SyncVP) model for RGB-D prediction consists of a latent diffusion model with two branches as illustrated in Fig. 2b. These branches are connected by a spatio-temporal cross-attention module placed between the deepest layers, right after the self-attention modules of the two denoising networks, allowing high-level semantic features alignment across the modalities. For each branch, we use a small custom version of the UNet architecture proposed by [56]. The latent autoencoder takes as input a sequence of frames $\mathbf{x} \in \mathbb{R}^{T \times H \times W \times C}$ and produces a latent vector $\mathcal{E}(\mathbf{x}) = \mathbf{z} = [\mathbf{z}^s, \mathbf{z}^h, \mathbf{z}^w] \in \mathbb{R}^{C' \times L}$, where $\mathbf{z}^s \in \mathbb{R}^{C' \times \frac{H}{4} \times \frac{W}{4}}$ encodes information about the general content of the video frames, while $\mathbf{z}^h \in \mathbb{R}^{C' \times T \times \frac{H}{4}}$ and $\mathbf{z}^w \in \mathbb{R}^{C' \times T \times \frac{W}{4}}$ encode temporal information. While one could define a simple cross-attention module that works on the intermediate features $\mathbf{z}_R = \mathcal{E}_R(\mathbf{r}_x)$ and $\mathbf{z}_D = \mathcal{E}_D(\mathbf{d}_x)$ of both UNets, we argue that a smarter and more efficient way would be to compute cross-attention only between the respective spatial and temporal latent feature vectors of the two modalities. Inspired by [23, 46], we propose a dual-way spatio-temporal cross-attention (STCA) that is efficiently computed by one single attention map. So given $\mathbf{z}_R = [\mathbf{z}_R^s, \mathbf{z}_R^h, \mathbf{z}_R^w]$ and $\mathbf{z}_D = [\mathbf{z}_D^s, \mathbf{z}_D^h, \mathbf{z}_D^w]$, the cross-attention is computed for each pair (3 times) of $\mathbf{z}_r \in \mathbf{z}_R$ and $\mathbf{z}_d \in \mathbf{z}_D$ as follows:

$$\mathbf{A} = \left(\frac{Q_R Q_D^\top}{\sqrt{d_k}}\right), \quad \begin{matrix} Q_R = \mathbf{W}_{Q_R}\mathbf{z}_r, & Q_D = \mathbf{W}_{Q_D}\mathbf{z}_d, \\ V_R = \mathbf{W}_{V_R}\mathbf{z}_r, & V_D = \mathbf{W}_{V_D}\mathbf{z}_d, \end{matrix}$$

$$\begin{aligned} \mathbf{z}_r &= \mathbf{z}_r + \mathbf{W}_{O_R}(\text{Softmax}\,(\mathbf{A})\,V_D), \\ \mathbf{z}_d &= \mathbf{z}_d + \mathbf{W}_{O_D}(\text{Softmax}\,(\mathbf{A}^\top)\,V_R), \end{aligned} \tag{1}$$

where $\mathbf{W}_{Q_R}, \mathbf{W}_{Q_D}, \mathbf{W}_{V_R}, \mathbf{W}_{V_D}, \mathbf{W}_{O_R}, \mathbf{W}_{O_D}$ are the query, key and output projection matrices for RGB and depth. Our split Spatio-Temporal Cross-Attention (STCA) not only results in better multi-modal video predictions than using normal cross-attention (CA) as we demonstrate in the experiments, but it is also more efficient. Indeed considering the dimension of our latent vector $\mathbf{z}$ and the quadratic cost of attention, the complexity ratio between STCA and CA is:

$$\frac{STCA(\mathbf{z})}{CA(\mathbf{z})} = \frac{O\left(\left(\frac{H \cdot W}{16}\right)^2 + \left(\frac{H \cdot T}{4}\right)^2 + \left(\frac{W \cdot T}{4}\right)^2\right)}{O\left(\left(\frac{H \cdot W}{16} + \frac{H \cdot T}{4} + \frac{W \cdot T}{4}\right)^2\right)}. \tag{2}$$

In case of 8 frames with resolution 64×64 or 128×128, STCA needs only 37% or 50% of the computation of CA, respectively. In our network, we implement STCA (1) as multi-head attention.

3.2. Cross-modality guidance

While the described architecture could be trained end-to-end to learn the joint distribution $p(\mathbf{r}_x, \mathbf{d}_x \mid \mathbf{r}_c, \mathbf{d}_c)$, such a straightforward training resulted to be ineffective. We argue that learning this distribution can be extremely complex due

to the differences between the two data modalities: RGB videos contain inherently more fine grained appearance and shading details compared to depth videos. A model trained from scratch on such data would struggle to converge as each modality has its own complexity and learning curve as we show in the experiments. To overcome this problem, we propose to first learn the two single conditional distributions $p(\mathbf{r}_x \mid \mathbf{r}_c)$ and $p(\mathbf{d}_x \mid \mathbf{d}_c)$ independently and then model the joint one in a second fine-tuning step. We therefore train two independent diffusion models based on the PVDM [56] UNet using the DDPM algorithm with the standard diffusion loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathcal{E}(\mathbf{x})_t, t, \mathcal{E}(\mathbf{c}))\|_2^2\right], \tag{3}$$

where θ are the parameters of the denoising model, $\mathbf{x}$ are the target future frames, t is the diffusion timestep and $\mathbf{c}$ are the conditioning frames. Since we apply diffusion in a latent space, we also need to train the autoencoders for each modality: $\mathcal{D}(\mathcal{E}(\cdot))_R$ for RGB and $\mathcal{D}(\mathcal{E}(\cdot))_D$ for depth. After an initial step of training independently the modality specific diffusion models ϵ_{θ_R} and ϵ_{θ_D}, we can use them as a reasonable starting point to model our joint conditional distribution $p(\mathbf{r}_x, \mathbf{d}_x \mid \mathbf{r}_c, \mathbf{d}_c)$. So the SyncVP branches, as shown in Fig. 2b, are initialized with these pre-trained weights and the multi-modal model is fine-tuned with respect to the following loss:

$$\mathcal{L}_M = \mathbb{E}_{\mathbf{r}_x,\mathbf{d}_x,\mathbf{r}_c,\mathbf{d}_c,\boldsymbol{\epsilon},t}\Big[\|\boldsymbol{\epsilon} - \epsilon_{\theta_R}(\mathcal{E}_R(\mathbf{r}_x)_t, t, \mathcal{E}_R(\mathbf{r}_c))\|_2^2 + \|\boldsymbol{\epsilon} - \epsilon_{\theta_D}(\mathcal{E}_D(\mathbf{d}_x)_t, t, \mathcal{E}_D(\mathbf{d}_c))\|_2^2\Big]. \tag{4}$$

Notice that, as illustrated in Fig. 2a, the target noise $\boldsymbol{\epsilon}$ is shared across both modalities. Specifically, we sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ only once and then apply the forward diffusion process to each modality's latent vector:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathcal{E}(\mathbf{x}_0) + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \tag{5}$$

where $\mathbf{x}_0$ are the original input frames and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, with $\alpha_t \in (0, 1)$, represents the noise schedule. In our experiments, we show that this is particularly beneficial for video prediction and we argue that since each modality is conditioned with initial frames that belong to a shared context, learning the same reverse denoising transformation simplifies the training and enforces the model to predict frames that are consistent across modalities, leading to faster convergence, lower loss, and better conditional generation.

Since our goal is to forecast multi-modal frames conditioned on both or only one modality as illustrated in Fig. 1, we propose a *cross-modality guidance* training procedure, which is inspired by the classifier-free guidance approach [15]. Instead of training the model only for the joint conditional distribution $p(\mathbf{r}_x, \mathbf{d}_x \mid \mathbf{r}_c, \mathbf{d}_c)$, we train our model for full and partial conditional generation, i.e., we simultaneously learn the following three distributions: $p(\mathbf{r}_x, \mathbf{d}_x \mid \mathbf{r}_c, \mathbf{d}_c)$, $p(\mathbf{r}_x, \mathbf{d}_x \mid \mathbf{0}, \mathbf{d}_c)$, $p(\mathbf{r}_x, \mathbf{d}_x \mid \mathbf{r}_c, \mathbf{0})$. This is achieved by randomly masking one of the modalities $\mathbf{r}_c$ and $\mathbf{d}_c$.

4. Experiments

4.1. Implementation details

Analyzing a single diffusion branch, we customize the PVDM model [56] as a two-level UNet, where channels are doubled, and the vector length is reduced by a factor of 4 only once. As a result, each modality-specific branch has approximately 58 million parameters, which is about 11% of the original PVDM-L model and 44% of the PVDM-S model. We train both single-modality models and SyncVP using DDPM with 1000 steps, and use DDIM [37] with 100 steps during inference. Cross-modality guidance is applied during training by conditioning on both modalities with a 50% probability, and on a single modality with a 25% probability each. Our model is trained to predict 8 frames in each forward pass and operates auto-regressively at test time to generate the desired sequence length. The source code is available at https://SyncVp.github.io/.

4.2. Datasets

We train and evaluate our multi-modal video prediction model on two widely used datasets for video prediction, Cityscapes [10] and BAIR [12], a subset of OpenDV-YouTube [51] at higher resolution, as well as two additional datasets, SYNTHIA [34] and ERA5-Land [31], to demonstrate the generalization of the method to other modalities beyond depth.

Cityscapes [10]: One of the most widely used datasets in video prediction benchmarks, Cityscapes provides RGB videos of driving scenes along with disparity maps computed from a stereo camera system mounted on a car. The dataset consists of short video sequences of 30 frames. Following previous works, we resize and center-crop the videos to a 128×128 resolution, using the first two frames as conditioning input to predict the remaining 28 frames ($2 \rightarrow 28$).

BAIR [12]: Known for its challenging stochasticity, BAIR contains over 40.000 videos of a robotic arm making highly random movements as it interacts with objects on a tabletop. BAIR provides only 64×64 RGB videos, each 30 frames in length, without depth information. To adapt this dataset to our setup, we compute pseudo ground-truth depth using DepthAnything-v2 [52]. Also in this case, following previous works, we adopt the $2 \rightarrow 28$ prediction setup.

OpenDV-YouTube [51]: The dataset comprises over 1,700 hours of real-world driving videos sourced from YouTube. For our experiments, we select a small subset approximately

Models	Cityscapes, 2 → 28			
	#T	FVD↓	SSIM↑	LPIPS↓
SVG-LP [11]	100	1300.26	0.574	549.0
NPVP [54]	100	768.04	0.744	183.2
VRNN 1L [7]	100	682.08	0.609	304.0
Hier-VRNN [7]	100	567.51	0.628	264.0
GHVAEs [47]	-	418.00	0.740	194.0
VDT [26]	-	142.3	**0.880**	-
MCVD [42]	10	141.31	0.690	112.0
ExtDM-K4 [59]	100	121.3	0.745	**108**
STDiff [55]	10	107.31	0.658	136.26
Ours (w/o depth)	10	97.31	0.652	161.1
Ours	10	**84**	0.649	159.7

Models	BAIR, 2 → 28			
	#T	FVD↓	SSIM↑	LPIPS↓
NPVP [54]	100	923.62	0.842	57.43
STM-FANet [19]	-	159.6	**0.844**	93.6
VPTR-NAR [53]	-	-	0.813	70.0
Hier-VRNN [7]	100	143.4	0.829	**55.0**
MCVD [42]	10	120.6	0.785	70.74
SAVP [22]	100	116.4	0.789	63.4
ExtDM-K4 [59]	100	102.8	0.814	69
STDiff [55]	10	88.1	0.818	69.40
Ours (w/o depth)	10	70.49	0.795	79.43
Ours	10	**63.60**	0.805	75.17

Table 1. Comparison of SyncVP with other methods on Cityscapes (left) and BAIR (right).

matching the size of Cityscapes (65 minutes at 30 fps). Frames are cropped and resized to 256×256, then these are grouped into non-overlapping clips of 32 frames. These are split into training (80%), validation (10%), and testing (10%) sets. We train this model in an $8 \rightarrow 8$ setting. Pseudo depth is computed using DepthAnything-v2 [52].

SYNTHIA [34]: This synthetic dataset of urban scenes provides frame-by-frame aligned RGB and semantic segmentation maps. SYNTHIA consists of 182 training and 90 test video clips, each with an average length of 500 frames. For our purposes, we use 16-frame video clips at a 128×128 resolution, training the model in an $8 \rightarrow 8$ setting, with the test set divided into smaller, non-overlapping clips of 16 frames each. We chose to use it to test our model for RGB and semantic maps, which is a different set of modalities.

ERA5-Land [31]: The dataset is widely used in global climate research, providing daily measurements from 1950 to the present as spatial images of size 360×540. In our experiments, we focus on two variables: the two-meter temperature (t2m) and surface pressure (sp). We restrict the dataset to data from 1979 onward, yielding a total of 16,799 frames. The raw reanalysis data are sourced from the Climate Data Store (CDS) [30]. Each frame is resized to 256×384, and the frames are then split into a training (80%) and testing (20%) set. The model is trained for $4 \rightarrow 4$ prediction and evaluated in a $4 \rightarrow 8$ setup.

4.3. Evaluation

Metrics For evaluation, we use Frechet Video Distance (FVD), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). FVD is the primary metric, as it evaluates both temporal coherence and perceptual quality by comparing distributions between real and generated videos in feature space. SSIM assesses spatial consistency frame-by-frame, focusing on luminance and structural attributes, while LPIPS measures perceptual similarity using deep features. Unlike SSIM and LPIPS, which ignore temporal dynamics, FVD captures high-level motion and is often considered to better align with human judgment, making it particularly suitable for evaluating video prediction tasks. Following the evaluation procedure of [42, 55], we also sample only 10 random future trajectories for each test sample and select the best one, as opposed to the 100 trajectories used by other methods. For a fair comparison, we report the number of trajectories (#T) used by previous methods whenever this information is clearly stated in the paper or the evaluation code is available. For depth and semantic segmentation, we report SSIM and L_2 error (multiplied by 100) between the generated frames and the ground-truth.

The ERA5-Land dataset evaluation is performed on 256 random samples from the test set, in this case we compute L_1 error for both sp and t2m.

Quantitative results We evaluate SyncVP with both RGB and depth conditioning frames on the Cityscapes and BAIR datasets in Tab. 1, demonstrating that our model surpasses previous state-of-the-art performance in FVD by over 21% and 27%, respectively. While SSIM and LPIPS are slightly lower than optimal, they remain comparable to previous approaches, particularly those using the same number of future trajectories per test sample (#T). We attribute the lower SSIM and LPIPS values to the significant compression applied by the autoencoder, as discussed in Sec. 3.1. Unlike prior methods that employ frame-by-frame autoencoders with a 3D latent representation, we use a spatio-temporal autoencoder that produces a simplified 2D latent vector. As shown in Tab. 2, this compact latent structure allows our approach to outperform previous diffusion models in inference speed, achieving faster results across both single and multi-modal scenarios.

Since we leverage depth as an additional informative cue

in the conditioning input, one might argue that our results are not directly comparable to prior methods that use only RGB. To address this, we also evaluate SyncVP with only RGB frames as conditioning (w/o depth) in Tab. 1, and demonstrate that it still achieves state-of-the-art performance. The first two rows in Fig. 6 show that it consistently generates depth frames aligned with the predicted RGBs despite of missing depth data in the input.

For SYNTHIA (Tab. 3) and ERA5-Land (Tab. 4), we show only a comparison between the single modality baselines and our multi-modal SyncVP, since this is the first work using these datasets for video prediction. In both cases, multi-modal training improves overall prediction performance. For ERA5-Land, surface pressure is measured in pascal (Pa) and two-meter temperature in kelvin (K).

	MCVD [42]	VDT [26]	ExtDM [59]	STDiff [55]	RGB only	SyncVP
Time (s)	37.72	24.34	30.31	239.4	**10.39**	22.68

Table 2. Average inference time comparison for predicting 28 frames with 2 conditioning frames at 128 × 128 resolution. We use 100 sampling steps and run the models on a NVIDIA TITAN RTX GPU with batch size 1.

Models	RGB			Sem. Segmentation	
	FVD↓	SSIM↑	LPIPS↓	SSIM↑	L_2 ↓
RGB	103.81	**0.827**	90.29	-	-
Sem. Segmentation	-	-	-	**0.649**	9.223
SyncVP	**93.37**	0.820	**89.92**	0.643	**8.733**

Table 3. Results on SYNTHIA (128 × 128, 8 → 8).

Models	sp (Pa) L_1 ↓	t2m (K) L_1 ↓
sp	462.69	-
t2m	-	1.85
SyncVP	**421.65**	**1.78**

Table 4. Results on ERA5-Land (256 × 384, 4 → 8).

Qualitative results We provide qualitative examples of SyncVP for the video prediction benchmarks BAIR and Cityscapes in Fig. 3 and Fig. 6, respectively. Fig. 3 shows an example on the BAIR dataset, where the random movements of the robotic arm are unpredictable after a few frames, but the RGB and depth predictions remain well-aligned throughout the sequence. Fig. 6 is particularly important to understand the effect of our cross-modality guidance training discussed in Sec. 3.2. Without our training strategy the model is not able to predict future frames for the missing modality (row 4).

Additional results on the SYNTHIA dataset are provided in Fig. 4, where the colorful predicted semantic segmentation maps allow to better appreciate the alignment between the two modalities. Fig. 7 shows predictions on climate data from the ERA5-Land dataset [31]. The results show the generalization of the method to other modalities. In Fig. 8, we provide some samples from the OpenDV-Youtube dataset [51].

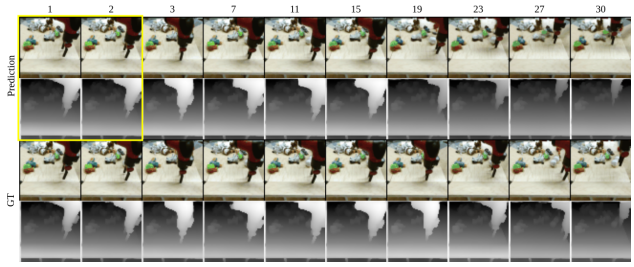


Figure 3. SyncVP predictions on BAIR using 2 conditioning frames (yellow frame). Predicting future movements of the robotic arm is challenging due to high stochasticity, but we can appreciate the alignment between predicted RGB and depth frames.

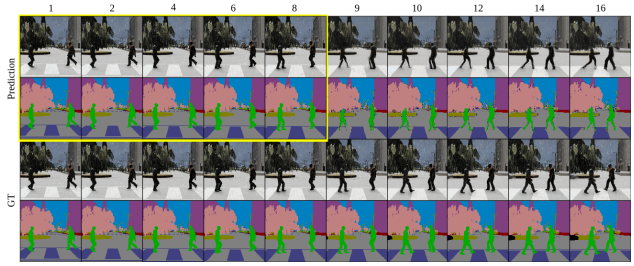


Figure 4. SyncVP predictions on SYNTHIA (RGB + Semantic segmentation) using 8 conditioning frames to predict the next 8 frames.

4.4. Ablations

To evaluate the effectiveness of our SyncVP model for multi-modal video prediction, we conduct several experiments on the Cityscapes dataset, comparing SyncVP's performance first with single-modality models and then with two other approaches for handling multi-modal information. Inspired by LDM3D [38], the [RGB,D] method concatenates the RGB and depth frames along the channel dimension and then trains a single autoencoder and diffusion model on them. In contrast, RGB+D with Fused Latents (FL), similar to [8], maintains separate autoencoders for each modality and fuses the latent representations by concatenating the latent vectors before applying a single diffusion model. Additionally, we assess the efficacy of our split Spatio-Temporal Cross-Attention module (STCA) compared to vanilla cross-attention mechanism.

All metrics for the models mentioned above are presented in Tab. 5 and have been computed using models trained for

the same amount of iterations: the RGB and depth models were trained for 1080k iterations, while the RGB + D (STCA) model was fine-tuned starting from the checkpoints at iteration 680k of the single-modality models for an additional 400k iterations. In addition to the notable improvements achieved by our SyncVP model, Fig. 5 illustrates how training with joint modalities significantly accelerates loss reduction compared to prolonged single-modality training. Next, we evaluate the impact of *cross-modality guidance* training. Specifically, Tab. 6 demonstrates the benefits of this approach on video prediction quality, while Fig. 6 shows that it enables the prediction of future frames for the missing modality, a skill that does not arise from a simple conditional training.

Finally, we assess the impact of using a shared forward diffusion process. As shown in Tab. 6, training the model with independent noise for each modality results in significantly lower performance across all metrics. Additional experiments done while exploring different design choices are reported in Tab. 7. The table shows that training from scratch, i.e. without pre-training the single modality models, performed poorly (row 1). Adding STCA to all layers (row 2) performed worse than adding it only to the latest layer (row 4). Adding a motion loss (row 3) that enforces similarity between temporal latent vectors of both modalities via an MLP projection resulted in a reduction in performance. We also investigated the differences between using shared or separate STCAs (rows 5-6). Since the performance was very similar, we chose the shared STCA as it is more efficient.

Models	RGB			Depth	
	FVD↓	SSIM↑	LPIPS↓	SSIM↑	L_2 ↓
RGB	142.51	0.659	173.64	-	-
Depth	-	-	-	0.825	8.000
[RGB,D]	123.8	0.662	184.21	0.812	8.161
RGB + D FL	160	**0.667**	180.82	**0.835**	7.401
RGB + D	97.68	0.652	162.87	0.829	7.459
RGB + D STCA	**84**	0.649	**159.73**	0.830	**7.329**

Table 5. Ablation on Cityscapes (128×128, $2 \rightarrow 28$). Comparing the performance of single modality baselines (rows 1 and 2) with multi-modal variants, using channel concatenation (row 3), single diffusion model with fused latents (row 4), and coupled diffusion models with vanilla (row 5) and split spatio-temporal cross attention (row 6).

5. Conclusions

In this paper, we introduce SyncVP, a novel versatile framework for multi-modal video prediction. This approach is the first to leverage informative non-RGB modalities videos while efficiently predicting all target modalities in a single pass. Built upon pre-trained diffusion models,

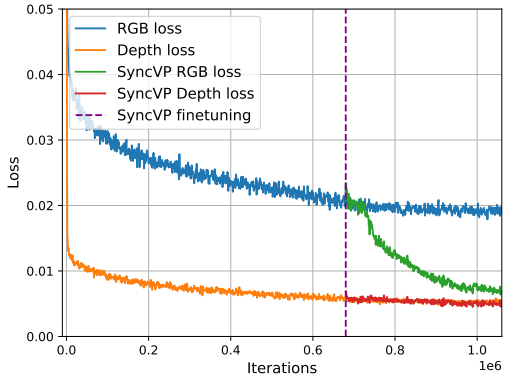


Figure 5. Comparison of RGB and depth loss for single-modality models and SyncVP on Cityscapes.

Same noise	Cross-modality guidance	RGB			Depth	
		FVD↓	SSIM↑	LPIPS↓	SSIM↑	L_2 ↓
✗	✓	143.16	**0.661**	173.64	0.826	8.122
✓	✗	122.4	0.657	169.22	0.828	7.525
✓	✓	**84**	0.649	**159.73**	**0.830**	**7.329**

Table 6. Ablation on Cityscapes about the impact of using the same noise vs. independent ones (rows 1 and 3), and the impact of cross-modality guidance (rows 2 and 3).

Variations	FVD↓	SSIM↑	LPIPS↓
Scratch	158.53	**0.674**	176.5
STCA at all layers	773.7	0.598	302.6
Motion loss	106.97	0.655	164.5
Ours	**84**	0.649	**159.7**
Non-shared STCA*	131.66	0.649	170.2
Ours*	129.92	0.650	171.1

Table 7. Impact of training and design choices on Cityscapes. * denotes only 100k iterations for training.

SyncVP employs a multi-branch diffusion network with spatio-temporal cross-attention to enable rich information exchange across modalities. As a result, SyncVP achieves new state-of-the-art performance on both the Cityscapes and BAIR benchmarks by over 21% and 27% on the main FVD metric respectively. Notably, our model demonstrates strong predictive capability even when one of the modalities is unavailable, suggesting that jointly generating multiple modalities inherently enhances the quality of the predicted frames. We believe that our approach can potentially pave the way to applications that require the fusion of diverse sensor inputs, a more comprehensive understanding of the predicted results and tolerance to eventual missing data.

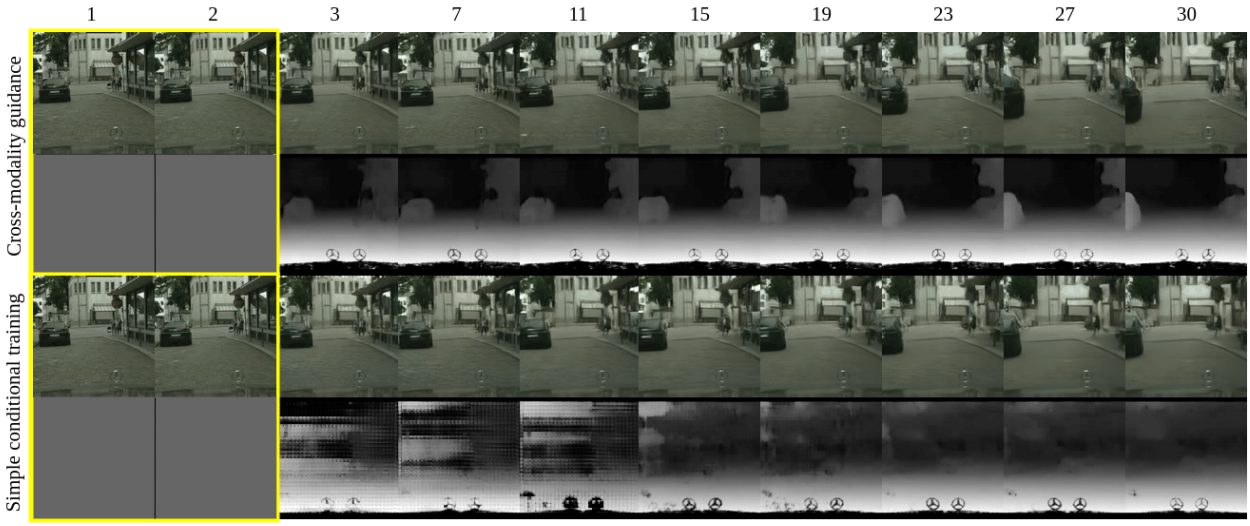


Figure 6. SyncVP predictions on Cityscapes using only RGB frames as conditioning. The third and fourth rows show the results where the model is trained without *cross-modality guidance*, which is crucial to predict the missing modality in future frames.

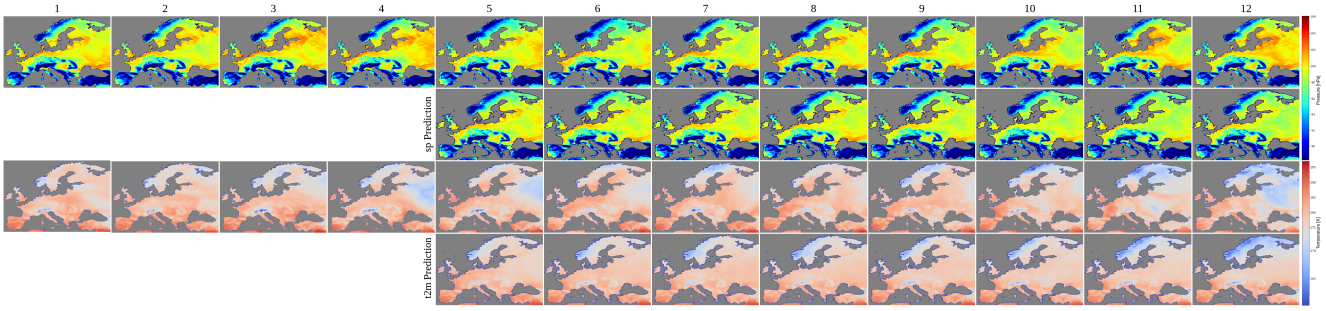


Figure 7. SyncVP predictions on ERA5-Land surface pressure (sp) and two-meter temperature (t2m) using 4 days measurements to predict the next 8 days.

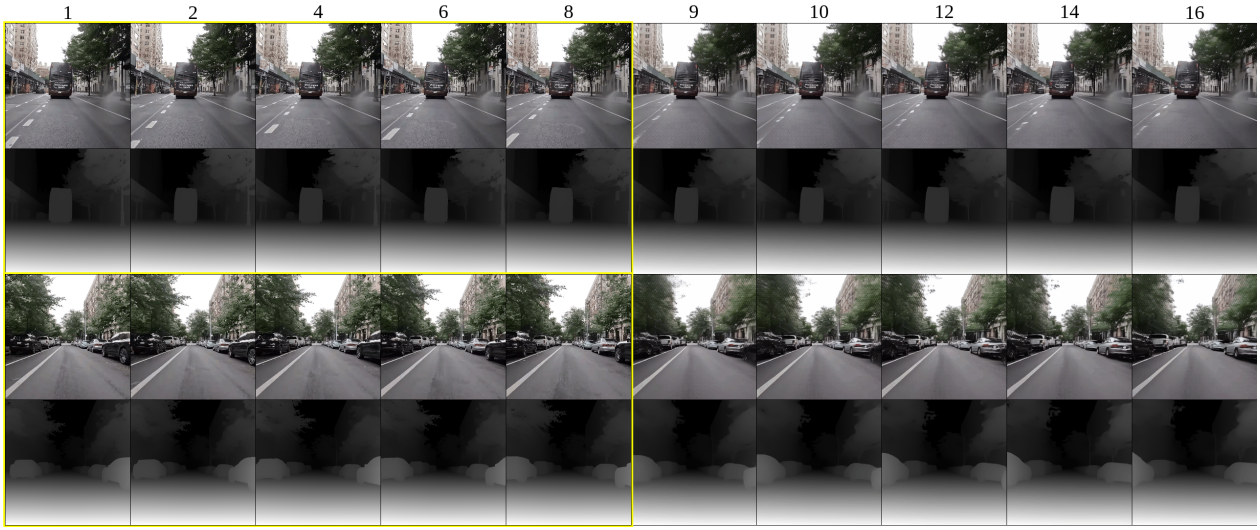


Figure 8. SyncVP predictions on OpenDV-Youtube [51].

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) GA 1927/9-1 (KI-FOR 5351), the Federal Ministry of Education and Research (BMBF) under grant no. 01IS24075C RAINA, and the ERC Consolidator Grant FORHUE (101044724). We sincerely acknowledge the EuroHPC Joint Undertaking for granting us access to Leonardo at CINECA, Italy, through the EuroHPC Regular Access Call (proposal No. EHPC-REG-2024R01-076). Additionally, the authors express their gratitude for the access provided to the Marvin cluster at the University of Bonn. We want to thank Mohamad Hakam Shams Eddin for preparing the ERA5-Land dataset.

References

[1] Adil Kaan Akan, Erkut Erdem, Aykut Erdem, and Fatma Güney. Slamp: Stochastic latent appearance and motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14728–14737, 2021. 2

[2] Adil Kaan Akan, Sadra Safadoust, and Fatma Güney. Stochastic video prediction with structure and motion. *arXiv preprint arXiv:2203.10528*, 2022.

[3] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018. 1

[4] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2020. 2

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023. 1, 2

[6] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. Contextvp: Fully context-aware video prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[7] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 5

[8] Changyou Chen, Han Ding, Bunyamin Sisman, Yi Xu, Ouye Xie, Benjamin Z. Yao, Son Dinh Tran, and Belinda Zeng. Diffusion models for multi-task generative modeling. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 6

[9] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019. 2

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[11] Emily L. Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, 2018. 2, 5

[12] Frederik Ebert, Chelsea Finn, Alex Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *Conference on Robot Learning (CoRL)*, 2017. 4

[13] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7346–7356, 2023. 1

[14] Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*. 4

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pages 8633–8646. Curran Associates, Inc., 2022. 2

[18] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *Transactions on Machine Learning Research*, 2022. 2

[19] B. Jin, Y. Hu, Q. Tang, J. Niu, Z. Shi, Y. Han, and X. Li. Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4553–4562, Los Alamitos, CA, USA, 2020. IEEE Computer Society. 2, 5

[20] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2024. 2

[21] Ariel Lapid, Idan Achituve, Lior Bracha, and Ethan Fetaya. Gd-vdm: Generated depth for better diffusion-based video generation. *ArXiv*, abs/2306.11173, 2023. 1

[22] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018. 2, 5

[23] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 3

[24] Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, pages 56–74. Springer, 2024. 1

[25] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[26] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. In *International Conference on Learning Representations*, 2023. 2, 5, 6

[27] Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan. Transformation-based adversarial video prediction on large-scale data. *arXiv preprint arXiv:2003.04035*, 2020. 2

[28] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10209–10218. IEEE Computer Society, 2023. 1

[29] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 1

[30] J. Muñoz Sabater. Era5-land hourly data from 1950 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, 2019. 5

[31] J. Muñoz Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, and S. Harrigan. Era5-land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, 2021. 4, 5, 6

[32] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. 1

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[34] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016. 4, 5

[35] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023. 1, 2

[36] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 4

[38] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv preprint arXiv:2305.10853*, 2023. 2, 6

[39] 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. 1, 2

[40] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *International Conference on Learning Representations*, 2017. 2

[41] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[42] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In *(NeurIPS) Advances in Neural Information Processing Systems*, 2022. 1, 2, 5, 6

[43] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[44] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017. 1, 2

[45] Zeyu Wang, Jingyu Lin, Yifei Qian, Yi Huang, Shicen Tian, Bosong Chai, Juncan Deng, Lan Du, Cunjian Chen, Yufei Guo, and Kejie Huang. Diffx: Guide your layout to cross-modal generative modeling. *ArXiv*, abs/2407.15488, 2024. 2

[46] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18909–18918, 2024. 3

[47] B. Wu, S. Nair, R. Martin-Martin, L. Fei-Fei, and C. Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2318–2328, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 2, 5

[48] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7623–7633, 2023. 1

[49] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yin-Yin He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Make-your-video: Customized video generation using textual and structural guidance. *IEEE transactions on visualization and computer graphics*, PP, 2023. 1

[50] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024. 1, 2

[51] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4, 6, 8

[52] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 4, 5

[53] Xi Ye and Guillaume-Alexandre Bilodeau. Vptr: Efficient transformers for video prediction. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3492–3499. IEEE, 2022. 2, 5

[54] Xi Ye and Guillaume-Alexandre Bilodeau. A unified model for continuous conditional video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3603–3612, 2023. 2, 5

[55] Xi Ye and Guillaume-Alexandre Bilodeau. Stdiff: Spatio-temporal diffusion for continuous stochastic video prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6666–6674, 2024. 2, 5, 6

[56] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 3, 4

[57] Yuanhao Zhai, Kevin Lin, Linjie Li, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, David Doermann, Junsong Yuan, Zicheng Liu, and Lijuan Wang. Idol: Unified dual-modal latent diffusion for human-centric joint video-depth generation. In *Proceedings of the European Conference on Computer Vision*, 2024. 1, 2

[58] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, XIAOPENG ZHANG, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[59] Zhicheng Zhang, Junyao Hu, Wentao Cheng, Danda Paudel, and Jufeng Yang. Extdm: Distribution extrapolation diffusion model for video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2024. 2, 5, 6