

# Cosmos-Transfer1: Conditional World Generation with Adaptive Multimodal Control

NVIDIA<sup>1</sup>

## Abstract

We introduce Cosmos-Transfer1, a conditional world generation model that can generate world simulations based on multiple spatial control inputs of various modalities such as segmentation, depth, and edge. In the design, the spatial conditional scheme is adaptive and customizable. It allows weighting different conditional inputs differently at different spatial locations. This enables highly controllable world generation and finds use in various world-to-world transfer use cases, including Sim2Real. We conduct extensive evaluations to analyze the proposed model and demonstrate its applications for Physical AI, including robotics Sim2Real and autonomous vehicle data enrichment. We further demonstrate an inference scaling strategy to achieve real-time world generation with an NVIDIA GB200 NVL72 rack. To help accelerate research development in the field, we open-source our models and code at <https://github.com/nvidia-cosmos/cosmos-transfer1>.

## 1. Introduction

Multimodal controllable world generation refers to the problem of generating world simulation videos based on multimodal video inputs such as segmentation, depth, and edge. These inputs help define specifics of the target world at different spatial locations at different time instances, which simplifies the generation problem. Such a controllable generation capability is valuable. One can leverage the capability to mitigate the synthetic-to-real domain gap problem of a CG-based simulator. We can make the renderings more realistic while preserving the scene structure and semantic through using depth and segmentation, often freely available in the CG-based simulator, as the multimodal inputs in world generation.

In this paper, we propose Cosmos-Transfer1, a diffusion-based conditional world model for the multimodal controllable world generation problem. Our model is built on top of Cosmos-Predict1 (NVIDIA, 2025), which consists of a set of diffusion transformer-based (DiT-based) (Peebles and Xie, 2023) world models. We add multimodal control branches to the DiT through a novel ControlNet design (Zhang et al., 2023). We build a control branch per modality. If there are  $N$  multimodal video inputs, then we will have  $N$  control branches. We train the  $N$  control branches separately and fuse them in the inference time.

Our multimodal control is spatially and temporally adaptive. This is achieved by applying a spatiotemporal control map to the outputs of the control branches. The spatiotemporal control map specifies the weight for each modality at each location and time instance. The higher the weight, the more influence the modality has on the generation output at the location and the time instance. This adaptive weighting scheme enables various ways to control the generation output. For example, one can choose to favor depth modality input more across all the locations in order to favor preservation of the input scene geometry. One can also give edge modality more weight to the foreground object to preserve the fine-grained details of the foreground object and less weight to the background to allow diverse background generation. The spatiotemporal control map could also be inferred by a separate neural module.

We conduct extensive empirical evaluations to verify the effectiveness of Cosmos-Transfer1. We measure its generation quality and controllability on several Physical AI related world generation tasks. We also discuss

<sup>1</sup>A detailed list of contributors and acknowledgments can be found in App. B of this paper.

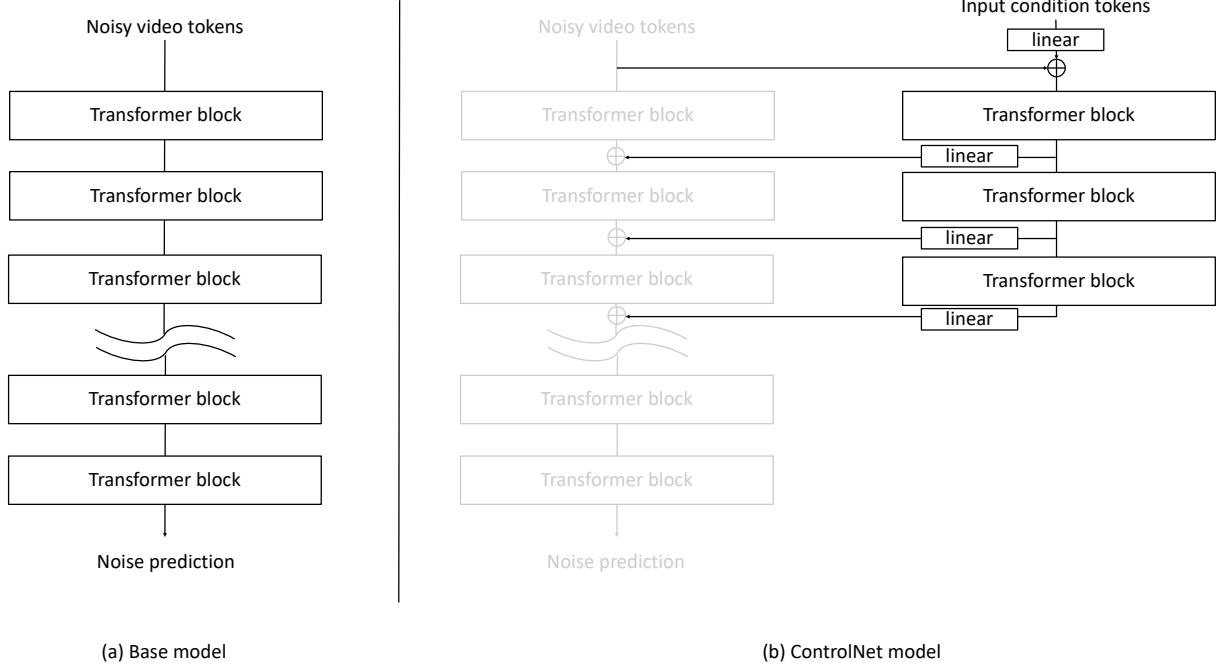


Figure 1: (a) **Base model** is the base DiT-based diffusion model. It consists of a sequence of transformer blocks and learns to predict the added noise in the input noisy tokens. (b) **ControlNet** extends the base model to a conditional diffusion model. The main addition is the control branch, which contains a few transformer blocks. The outputs of the transformer blocks are passed to zero-initialized linear layers before added back to the main branch. During the ControlNet training, the base model weights are frozen.

applications of Cosmos-Transfer1 for robotics Sim2Real and autonomous vehicle data enrichment. We further demonstrate an inference scaling strategy to achieve real-time world generation with an NVIDIA GB200 NVL72 rack. To help advance the field, our code, model weights, and example scripts are open-sourced at <https://github.com/nvidia-cosmos/cosmos-transfer1>.

## 2. Preliminary

The main component of a diffusion model is a denoiser. A common approach to implement the denoiser is to use the DiT architecture as visualized in Fig. 1(a). The architecture consists of a sequence of transformer blocks, trained to predict the noise  $\mathbf{b}$  added to the input video tokens. Let  $D$  denote the denoiser. In the diffusion model, it takes noisy video tokens  $\mathbf{x}_\sigma$  and noise deviation  $\sigma$  as inputs and uses them to predict the noise. That is

$$\mathbf{n} = D(\mathbf{x}_\sigma, \sigma). \quad (1)$$

ControlNet is a popular approach to extend the base diffusion model to a conditional diffusion model. The original ControlNet was developed for the UNet model. Chen et al. (2024) extended it to the DiT architecture. Based on the prior works, we design our DiT-based ControlNet visualized in Fig. 1(b). It has the base model and a control branch. The base model is the base diffusion model. The control branch consists of a few transformer blocks for conditional inputs. These conditional inputs serve as the control signal. In our ControlNet design, we use three conditional blocks. We choose three as it empirically offers a good balance between control effectiveness and inference efficiency. These transformer blocks are initialized by inheriting the weights from the corresponding transformer blocks in the base diffusion model. Outputs from these blocks are passed to corresponding linear layers, respectively. These linear layer outputs are then added to the activations of the corresponding transformer block in the base model. During training, these

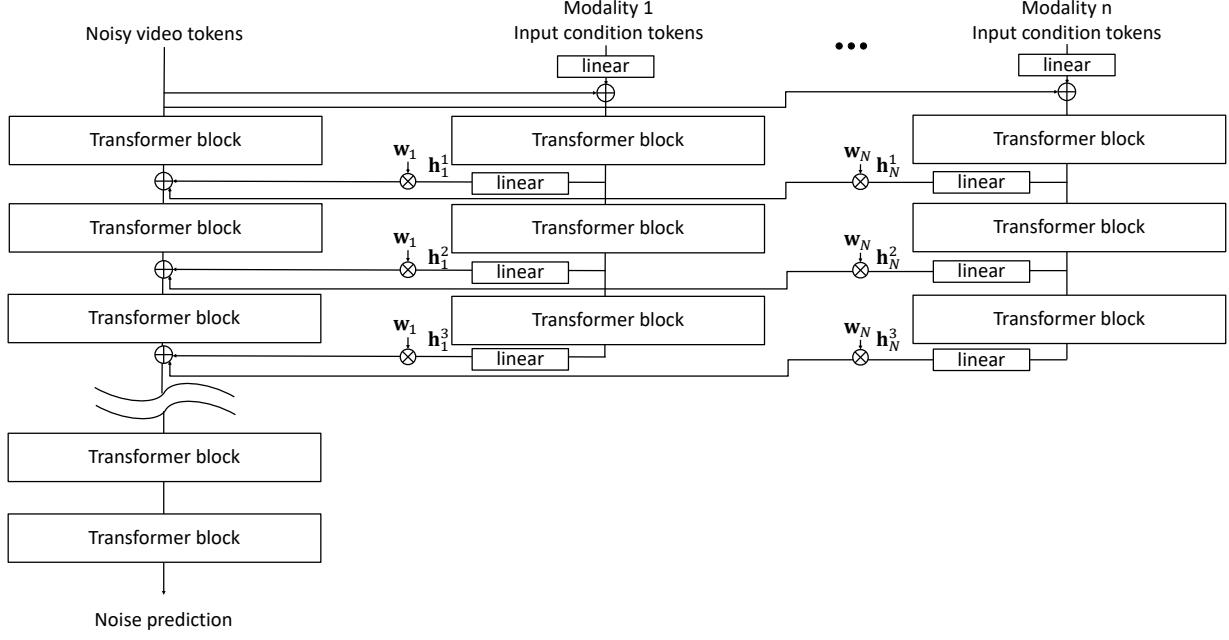


Figure 2: **Cosmos-Transfer1** is a world generator with adaptive multimodal control. It contains multiple control branches to extract control information from different modality inputs such as segmentation, depth, and edge. We apply spatiotemporal control maps  $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N\}$  to weight the outputs computed by different control branches before channeling them back to the main generation branch. The spatiotemporal control map allows the model to leverage the most relevant modalities in different regions for optimal output quality.

additional blocks are optimized while the base model is kept frozen. A well-trained denoiser learns to leverage information in the conditional tokens to predict the noise. Let  $\mathbf{c}$  be the conditional tokens. The denoiser for the ControlNet becomes a conditional denoiser. It predicts  $\mathbf{n}$  based on  $\mathbf{x}_\sigma$ ,  $\sigma$ , and  $\mathbf{c}$ .

$$\mathbf{n} = D(\mathbf{x}_\sigma, \sigma, \mathbf{c}). \quad (2)$$

We also note that ControlNet inherits the base model capability. When the base diffusion model takes text prompts as inputs, the ControlNet also takes text prompts as inputs.

### 3. Method

The proposed Cosmos-Transfer1 is a diffusion-based multimodal controllable world generator. It is constructed by post-training the Cosmos-Predict1 diffusion world model ([NVIDIA, 2025](#)). Let  $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_N$  be  $N$  conditional inputs, representing  $N$  different modalities. Cosmos-Transfer1 learns to leverage these  $N$  input videos to generate the world simulation.

For additional control, we further introduce the spatiotemporal control map  $\mathbf{w} \in \mathbb{R}^{N \times X \times Y \times T}$  where  $X$ ,  $Y$ , and  $T$  are the video width, height, and frame counts. The spatiotemporal control map enables fine-grained, adaptive control, allowing the model to leverage the most relevant modalities in different spatiotemporal regions for optimal output quality. Let  $\mathbf{h}_i^j \in \mathbb{R}^{X \times Y \times T}$  be the activations from the  $j$ th block of the  $i$ th control branch. Also, let  $\mathbf{w}_i \in \mathbb{R}^{X \times Y \times T}$  be the  $i$ th slice of the spatiotemporal control weight. The final activation from the  $j$ th block of the  $i$ th control branch to be added back to the main branch is given by  $\mathbf{w}_i \cdot \mathbf{h}_i^j$  where  $\cdot$  denotes the element-wise product. We visualize our adaptive multimodal ControlNet in Fig. 2.

The spatiotemporal control maps can be derived through various approaches. First, they can be manually designed. Second, they can be adaptively derived from heuristic rules based on prior observations of the

modalities. Lastly, one can train a neural module to predict the spatiotemporal control map. Note that at each spatiotemporal site, if the sum of the control maps across different modalities is greater than one, we apply normalization to the modality weights so that they sum up to one.

We train our individual control branches separately following the practice described in Sec. 2 and fuse them only at inference. This strategy has several advantages as compared to directly training all the control branches at once. First, it is more memory efficient, as we only need to fit one control branch into memory at a time during training. This divide-and-conquer approach helps ease the implementation burden in large-scale world model training where memory-expensive video data are used. Second, different modalities can be trained with different data, as paired data may be difficult to obtain for certain modalities. Finally, it is more flexible, as we can easily add or remove modalities at inference time after the training is done.

## 4. Modality and Training

Our first realization of Cosmos-Transfer1 is a post-trained Cosmos-Predict1-7B-Video2World model (NVIDIA, 2025), referred to as Cosmos-Transfer1-7B. We also realize a special version of Cosmos-Transfer1 for autonomous vehicle tasks. The model is post-trained from Cosmos-Predict1-7B-Video2World-Sample-AV model (NVIDIA, 2025), which is a finetuned version of Cosmos-Predict1-7B-Video2World on dash cam videos. We will term this model Cosmos-Transfer1-7B-Sample-AV.

For Cosmos-Transfer1-7B, we train the control branches with the high-quality finetuning dataset in (NVIDIA, 2025). For ease of reference, when Cosmos-Transfer1-7B operates on the single modality setting—the basic ControlNet setting—we will refer to it as Cosmos-Transfer1-7B [modality name], such as Cosmos-Transfer1-7B [Depth]. We train each control branch with 1024 NVIDIA H100 GPUs for a period of 2 to 4 weeks depending on the modality. Inherited from Cosmos-Predict1-7B-Video2World, an inference call of Cosmos-Transfer1-7B generates a 5-second 1280x704p video under 24 fps, which translates to 56K tokens<sup>2</sup> predicted by the model.

We introduce the following modalities in the paper:

- **Blur visual.** We apply bilateral blur (Tomasi and Manduchi, 1998) to the input video to create a blurry video as the input modality. This control is useful when we plan to keep the original colors and rough shapes while only adding or changing texture details of the input. For example, it can bring realistic edge transition while improving details and clarity when we transfer CG-rendered videos to realistic videos. During training, we randomize various parameters in bilateral blur as a data augmentation strategy. We will call Cosmos-Transfer1-7B operating in this setting Cosmos-Transfer1-7B [Vis].
- **Edge.** We extract Canny edges (Canny, 1986) from the input video frame-by-frame, and feed the edge video as the input modality to the control branch. This modality is useful when we only want to keep the overall structure of the scene, but give the model more freedom to creatively fill in the rest. During training, we randomize various thresholds in the Canny edge detector as a data augmentation strategy. We will call Cosmos-Transfer1-7B operating in this setting Cosmos-Transfer1-7B [Edge].
- **Depth.** We compute depth maps from the input video using DepthAnything2 (Yang et al., 2024) and then normalize the extracted depth to [0, 1] to generate a depth video. This control is particularly useful when keeping the 3D geometry of the input world is essential. We will call Cosmos-Transfer1-7B operating in this setting Cosmos-Transfer1-7B [Depth].
- **Segmentation.** We use GroundingDino (Liu et al., 2024) to detect objects in the caption from the first frame of the video. We then use SAM2 (Ravi et al., 2024) to extract object segmentation masks from the whole video. Since there can be infinitely many object categories, we randomize the colors in the segmentation masks, *i.e.* colors no longer have semantic meanings and only represent different objects. This model is advantageous when we aim to keep the original segmentation layout of the scene while

<sup>2</sup>We use Cosmos-Tokenize1-CV8x8x8-720p, which is a causal video tokenizer with 8x8x8 compression rate across t-axis, x-axis, and y-axis (NVIDIA, 2025). 56,320 tokens is computed as: 1280 (width)  $\div 8$  (tokenize)  $\div 2$  (patchify)  $\times$  704 (height)  $\div 8$  (tokenize)  $\div 2$  (patchify)  $\times$  [(121 - 1)  $\div 8 + 1$ ] (tokenized frames).

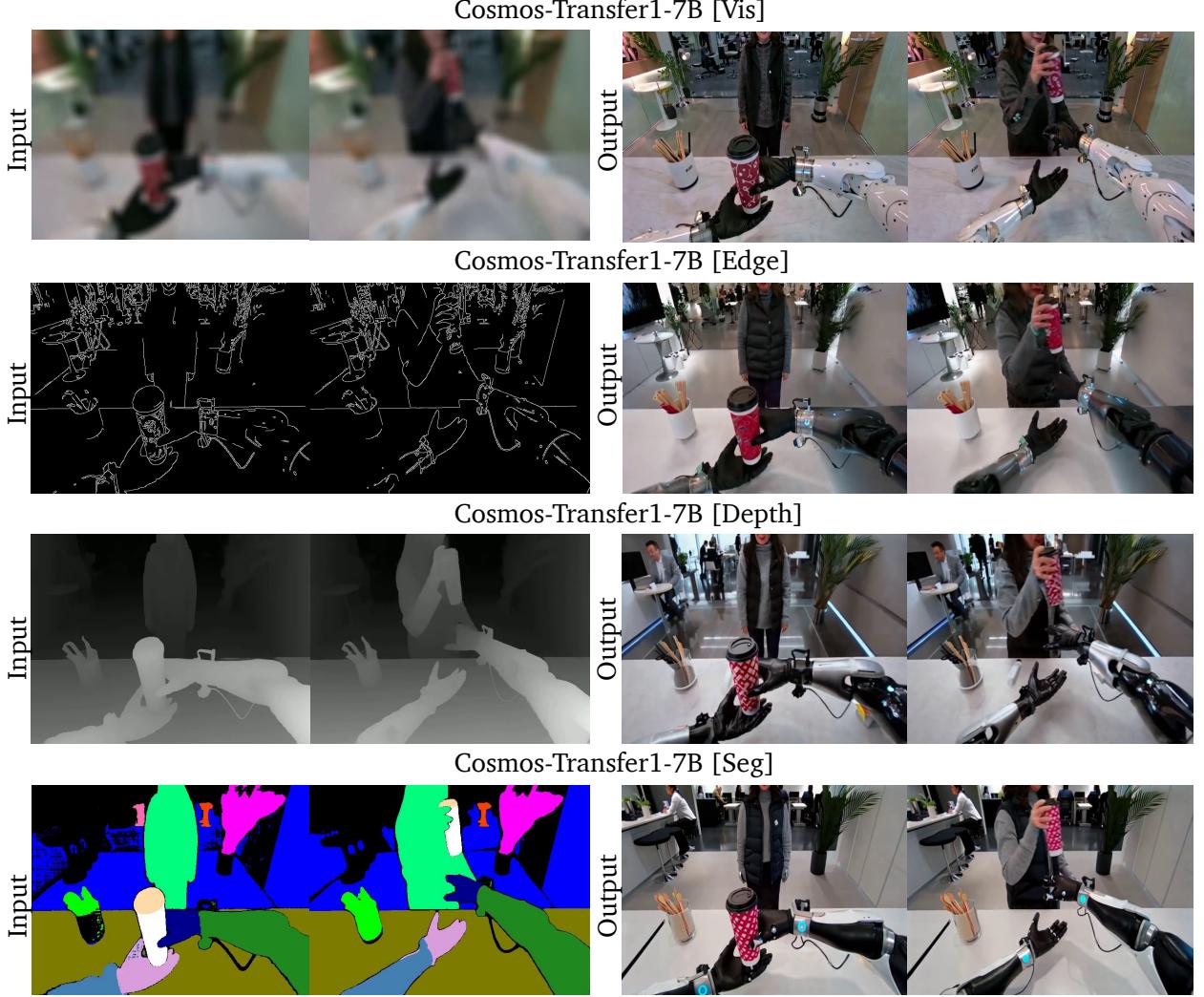


Figure 3: **Input and generated videos from Cosmos-Transfer1-7B operating on individual modality settings using the same prompt.** In particular, Cosmos-Transfer1-7B [Vis] preserves the colors and overall composition while altering texture details. On the other hand, Cosmos-Transfer1-7B [Edge] maintains the object boundaries while changing colors. Similarly, Cosmos-Transfer1-7B [Depth] preserves the scene geometry, while Cosmos-Transfer1-7B [Seg] preserves the scene semantics.

generating a new video with a high degree of freedom. We will call Cosmos-Transfer1-7B operating in this setting Cosmos-Transfer1-7B [Seg].

We use a different dataset for training Cosmos-Transfer1-7B-Sample-AV. In autonomous driving applications, HD maps with dynamic 3D bounding boxes are commonly used perception signals, while LiDAR scans are typically collected as a complement to RGB videos. Together, these modalities provide a comprehensive semantic representation of a 3D environment in the autonomous vehicle setting. Based on this motivation, we curated a 360-hour, high-quality autonomous driving dataset with additional HD map and 3D bounding box annotations. We call this newly curated high-quality dataset Real Driving Scene HQ (RDS-HQ). RDS-HQ comprises of 65K 20-second surrounding-view video clips (equivalent to approximately 360 hours of videos) along with corresponding 10 Hz LiDAR scans, captured with the NVIDIA driving platform. Each clip is annotated with dense captions to enhance the alignment with domain-specific details like camera mounting position and

contender vehicle density.

The modalities used by Cosmos-Transfer1-7B-Sample-AV are listed below.

- **HDMap.** High-definition (HD) maps provide detailed and precise road annotations, including lane lines, road boundaries, stop lines, poles, crosswalks, road markings, traffic lights, and traffic signs. We constructed our HD map using a pre-built city-level LiDAR map, leveraging an in-house auto-labeling pipeline. We further augmented the HD map annotations with 3D bounding box detection and tracking of dynamic objects, including vehicles, pedestrians, and other vulnerable road users. All auto-labeled annotations undergo rigorous human quality assessment (QA) to ensure high accuracy. The HDMap ControlNet is particularly advantageous when preserving the original road layout of a driving scene while generating a new video with a high degree of freedom. Additionally, it plays a crucial role in simulating different ego-car trajectories, as HD maps provide precise information on the relative pose translation between the ego camera and the road geometry. We will call Cosmos-Transfer1-7B-Sample-AV operates in this setting Cosmos-Transfer1-7B-Sample-AV [HDMap].
- **LiDAR.** To synchronize the collected 10 FPS LiDAR scans with the 30 FPS camera frames, we temporally interpolate and densify the LiDAR data. For each camera frame, we select the nearest LiDAR scan along with four additional adjacent scans (two preceding and two subsequent) to provide enhanced temporal context. Static points from these LiDAR scans are directly projected onto the camera frame, while for the dynamic points, identified using the object detection auto-labeling pipeline, we adjust their positions by interpolating the 3D bounding boxes prior to projection. To further reduce projection-induced holes, we apply a kernel-based interpolation with a kernel size of 4. The LiDAR ControlNet can be utilized to preserve the original semantic details of the driving scene while allowing modifications to environmental conditions and lighting. We will call Cosmos-Transfer1-7B-Sample-AV operates in this setting Cosmos-Transfer1-7B-Sample-AV [LiDAR].

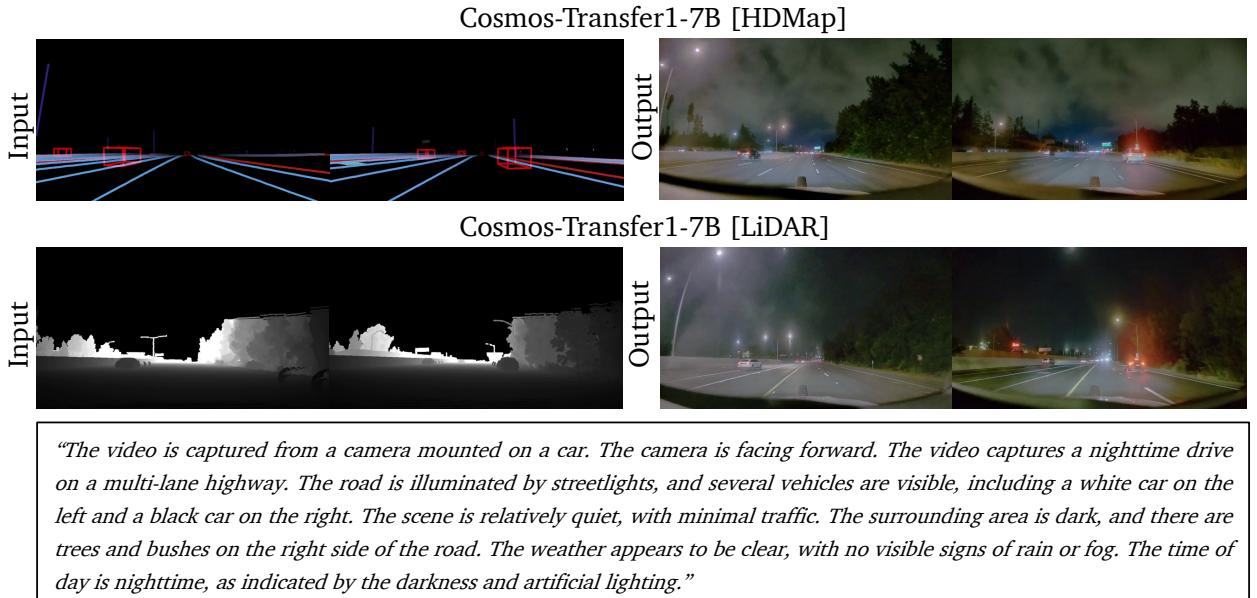


Figure 4: **Input and generated videos from Cosmos-Transfer1-7B-Sample-AV operating on individual modality settings.** Cosmos-Transfer1-7B-Sample-AV [HDMap] preserves the original road layout of a driving scene while Cosmos-Transfer1-7B-Sample-AV [LiDAR] preserves the input semantic details.

Finally, we also train an Upscale ControlNet to upscale the generated videos from 720p to 4k resolution. We show examples of the 4KUpscaler in Fig. 5. We take random crops of corrupted high-resolution videos using the corruption techniques in Real-ESRGAN (Wang et al., 2021) as input. We then train the ControlNet to recover

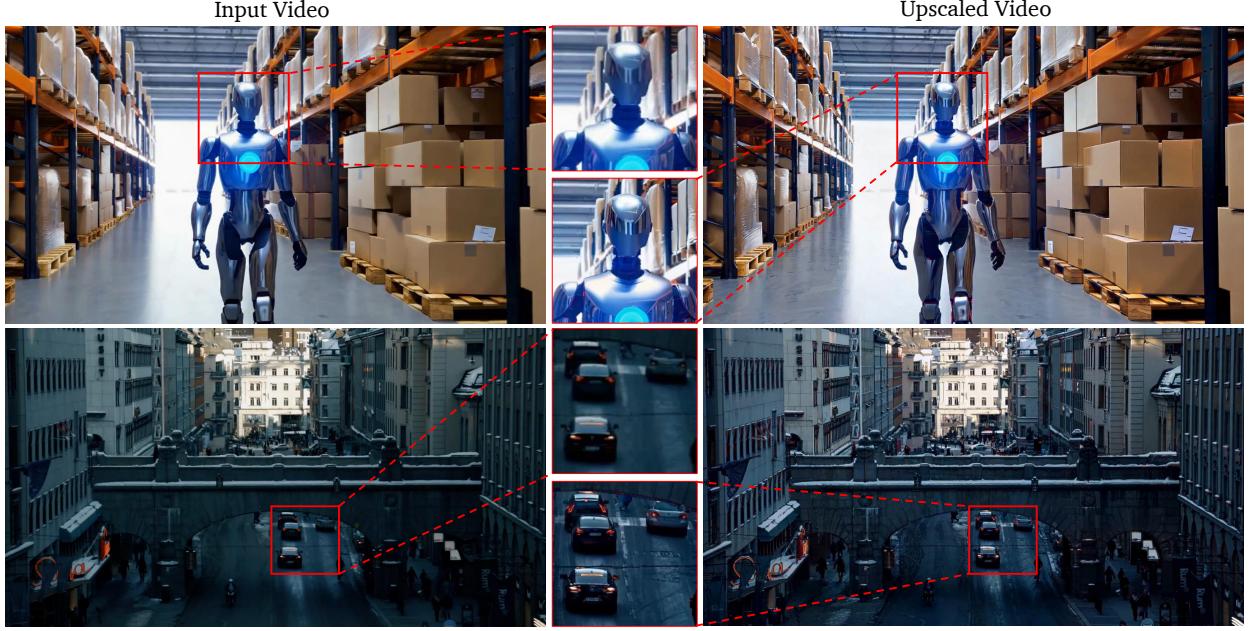


Figure 5: **Cosmos-Transfer1-7B-4KUpScaler upscaling videos from 720p to 4k resolution.** The input video in the first row is a generated video, while the second row is a real video. Note how the model adds realistic reflections and sharpens the textures in the input.

the original high-quality video patches. During inference, we adopt patch-based generation and divide the 4k output into  $3 \times 3$  grids with overlapping regions. At each denoising step, we run inference for each grid and average their results in the overlapping regions. This ensures the output video is seamless around the boundaries. We will call the 4KUpScaler version Cosmos-Transfer1-7B-4KUpScaler.

## 5. Evaluations

In this section, we present a comprehensive evaluation of Cosmos-Transfer1. We begin by introducing the dataset and evaluation metrics. In Sec. 5.1, we first show basic characteristics of the models by comparing single control models with multimodal control models that use uniform spatial weights. In Sec. 5.2, we present an approach to automatically create spatiotemporal control maps, and analyze the characteristics of Cosmos-Transfer1 when utilizing them. Finally, Sections 5.3 and 5.4 provide case studies demonstrating the applicability of our method for Sim2Real data generation in robotics and autonomous driving, respectively.

**TransferBench.** To evaluate the characteristics of Cosmos-Transfer1, we curate *TransferBench* — an evaluation dataset consisting of 600 examples across three key scenarios: robotic arm operations, driving, and ego-centric everyday life scenes, each representing a critical aspect of Physical AI. Specifically, we randomly sample 200 examples from the **AgiBot World** dataset (AgiBot-World-Contributors et al., 2025) for robotic arm operations, which capture fine-grained manipulation, dexterity, and interaction with objects, essential for automation and industrial robotics. For driving scenarios, we sample 200 examples from the **OpenDV** dataset (Yang et al., 2024), representing mobile autonomy, perception in dynamic environments, and decision-making in complex traffic conditions. Lastly, for ego-centric everyday life scenes, we sample 200 examples from the **Ego-Exo-4D** dataset (Grauman et al., 2024), focusing on human-centric AI, interaction modeling, and embodied perception in unstructured settings. This selection provides a diverse evaluation suite, covering both structured and unstructured real-world environments that are handy for benchmarking Cosmos-Transfer1 performance.

**Metrics.** We evaluate Cosmos-Transfer1 based on three key aspects: 1) Adherence to Control Input Signals, 2) Generation Diversity, and 3) Overall Generation Quality.

- **Adherence to Control Input Signals.** For quantitative evaluation, we first transform both the input sample videos and the generated videos into a shared representation space by applying operations, including blurring, edge extraction, depth estimation, and semantic segmentation. We then compare these transformed representations to measure alignment, assessing whether the generated worlds accurately reflect the intended conditioning. Specifically, for the alignment with each of the control signals, we perform the following:
  - **Vis Alignment.** We apply the same blurring operation to both the input sample videos and the generated videos, and compute their Structural Similarity Index Measure (SSIM) (Wang et al., 2004). We then average the scores over all the samples within the dataset. We call this metric **Blur SSIM**. Higher values mean better alignment.
  - **Edge Alignment.** We apply the same Canny edge extraction to both the input sample videos and the generated videos, and compute a classification F1 score (van Rijsbergen, 1979) on the black-and-white pixel classification. We then average the F1 scores over the dataset. We call this metric **Edge F1**. Higher values mean better alignment.
  - **Depth Alignment.** We compute the scale-invariant Root Mean Squared Error (si-RMSE) (Eigen et al., 2014) between the depth maps extracted from both the input sample videos and the generated videos using DepthAnythingV2 (Yang et al., 2024). We call this metric **Depth si-RMSE**. Lower values mean better alignment.
  - **Segmentation Alignment.** We compute the mean Intersection over Union (mIoU) between the segmentation masks generated using our GroundingDINO+SAM2 annotation pipeline (Ravi et al., 2024). Since GroundingDINO is an open-set object detection model, it might detect the same object multiple times, resulting in repeated instance segmentations. To address this, we aggregate all segmentation masks associated with the same phrase in the captions into a single segmentation mask, ensuring comparisons are made at the object level across the video. Next, we establish correspondences between the ground truth and generated segmentation masks using an matching algorithm based on the IoU distance function. Masks with an IoU below 0.1 are discarded to eliminate spurious matches, and the final mIoU score is computed as the mean IoU over all valid correspondences. We call this metric **Mask mIoU**. Higher values mean better alignment.
- **Diversity.** For a given condition input, we expect Cosmos-Transfer1 can generate different videos based on different text prompts. To evaluate diversity, we design the following metric based on LPIPS (Zhang et al., 2018). For each condition input, we generate  $K$  videos from  $K$  different text prompts. We compute the LPIPS scores for the  $K(K - 1)/2$  pairs of the  $K$  videos. We average the LPIPS scores over all these pairs to get the averaged perceptual similarity. We then average over all samples in the dataset. We call this metric **Diversity-LPIPS**. Higher values mean better diversity.
- **Overall Generation Quality.** We report the DOVER-technical score (Wu et al., 2023), which serves as a perceptual metric for assessing the general aesthetic quality of the generated videos. A higher DOVER score indicates better visual quality. We average the DOVER scores across all videos within the dataset. We call this metric **Quality Score**.

### 5.1. Unimodal versus Multimodal

Before introducing the spatiotemporal control map, we first characterize the impact of different control modalities by evaluating several special cases of Cosmos-Transfer1:

- *Single control model.* In the single-control setting, the model is conditioned on only one control modality: blurred original video, Canny edges, depth maps, or segmentation maps, while the others are disabled. These settings help understand the strength of individual modalities.
- *Multimodal control model with one single modality excluded.* In this setting, except one modality, all other modalities contribute equally. These settings help understand the weakness of the overall model when one modality is excluded.

- *Full multimodal control model with uniform weights.* In this setting, we fuse all the modalities with an equal weight across the spatiotemporal control map.

[Tab. 1](#) shows the experiment results. As expected, the vis control model achieves the highest Blur SSIM (0.96), reflecting its strength in preserving coarse structure and color. Similarly, the edge control model attains the best Edge F1 score (0.28), underscoring its ability to capture dense structural details in the control input<sup>3</sup>. The depth control and segmentation control models do not achieve the highest scores in their respective metrics; the depth control model records a Depth si-RMSE of 0.49 (comparable to the blur visual control model), while the segmentation control model achieves a Mask mIoU of 0.68. As shown in [Fig. 3](#) and [Fig. 6](#), depth maps can be largely homogeneous in distant background regions, while segmentation maps from SAM2 often comprise broad color regions rather than fine-grained structures. This sparsity imposes fewer constraints on the generation process, giving the model additional freedom that may lead to less accurate reconstructions when these modalities are used in isolation.

Similar patterns can be observed when we only exclude a single modality and assign uniform weights to other modalities. For example, when only blur visual control is excluded (*Cosmos-Transfer1-7B Uniform Weights, no Vis*), it achieves the lowest Blur SSIM (0.68) compared to others, indicating that the dense structural cues provided by the blur visual input are critical for high alignment. Notably, the diversity scores reveal that modalities with denser structural information (Blur visual and Edge) tend to yield lower diversity, whereas those with sparser cues (Depth and Segmentation) promote higher diversity. Specifically, excluding Blur visual or Edge increases the diversity LPIPS to 0.37 and 0.31, respectively, while removing Depth or Segmentation reduces it to 0.25 and 0.23, respectively. The above observations suggest that the Blur visual and Edge controls, by providing dense structural information, are particularly useful for tasks requiring precise alignment and high-fidelity output—such as photorealistic style transfer or Sim2Real transformation in controlled environments. In contrast, Depth and Segmentation controls, which offer sparser structural guidance, allow for greater output variability and are thus suited for applications where diversity is desired, such as novel scene synthesis for data augmentation.

In contrast, the adaptive multimodal control model (*Cosmos-Transfer1-7B Uniform Weights*), which leverages multiple control inputs, consistently produces high-quality results. Although it ranks second in both Vis and Edge Alignment—likely because each modality is assigned a lower weight (0.25) compared to 1.0 in the single-control setting, it achieves the best depth reconstruction and the highest Quality Score (8.54). This suggests that, while individual control inputs excel at capturing specific aspects of the video, our multimodal control strategy effectively integrates complementary features from all modalities, leading to a more balanced and accurate output. These results validate our motivation for the multimodal control design of Cosmos-Transfer1. In the following sections, we introduce a method for automatically generating spatially-varying control weights and analyze how Cosmos-Transfer1 behaves when applying region-specific control.

## 5.2. Case Study for Spatiotemporal Control Maps

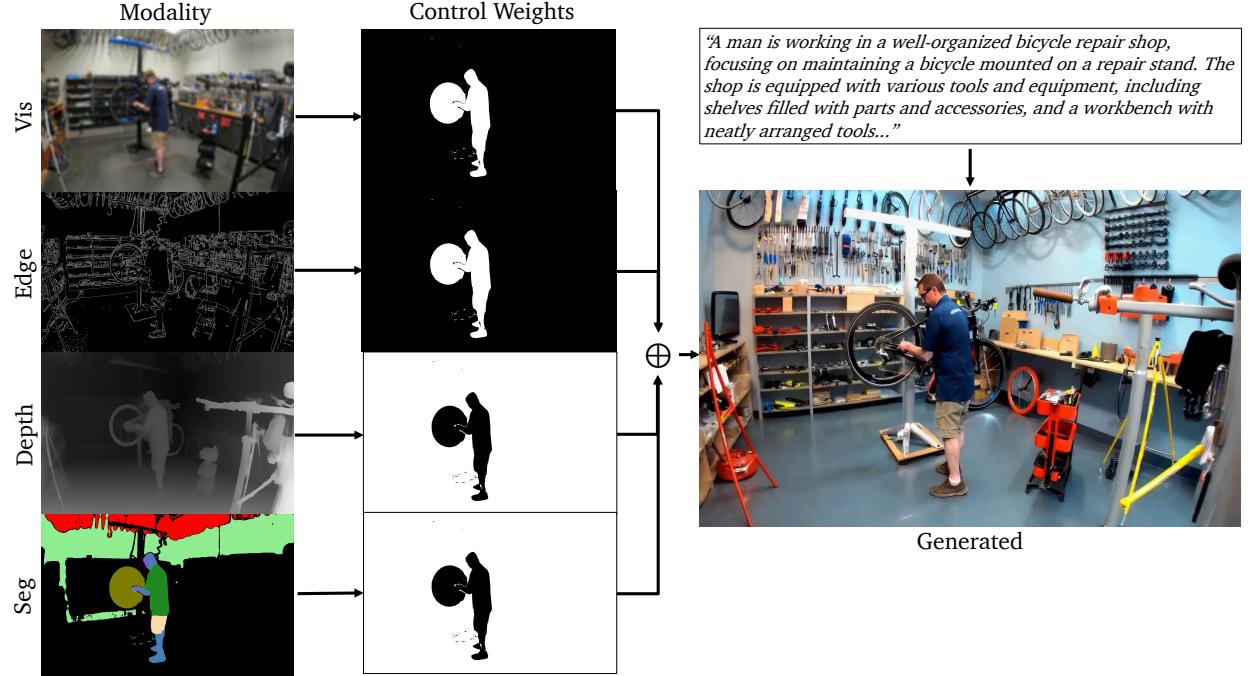
To demonstrate the strength of the proposed spatiotemporal control map, we design a *SalientObject* algorithm where we give different modalities different weights based on whether the location is from foreground or background. Specifically, we prompt a VLM ([Achiam et al., 2024](#)) to classify each pre-extracted GroundingDINO+SAM2 mask into either foreground and background. To do so, we provide the model with (1) reference video frames, (2) the complete caption we use for generation, and (3) the list of phrases associated to each segment mask, and then prompt the model to classify whether each phrase belongs to a salient foreground object in the scene.

Following this procedure, we can create a spatiotemporal foreground-background control weight mask. We show an example in [Fig. 6](#), where we assign a weight of 0.5 to vis and edge modalities respectively in the foreground,

<sup>3</sup>The Edge F1 score is a very strict metric that requires pixel-level alignment of the extracted thin canny edges. This strictness contributes to relatively low F1 scores for all methods.

**Table 1: Quantitative evaluation on TransferBench for various Cosmos-Transfer1 configurations.** We compare single control models (each conditioned on a single modality) with multimodal variants that use spatially uniform weights. For the multimodal cases, “Cosmos-Transfer1-7B, Uniform Weights” denotes the full model that integrates all four control modalities (each weighted at 0.25), while variants such as “Cosmos-Transfer1-7B, Uniform Weights, No Vis” exclude a specific modality (here, the blur visual control), with the remaining modalities retaining equal weights. Best results are in bold; second-best are underlined.

Model	Vis Alignment	Edge Alignment	Depth Alignment	Segmentation Alignment	Diversity	Overall Quality
	Blur SSIM $\uparrow$	Edge F1 $\uparrow$	Depth si-RMSE $\downarrow$	Mask mIoU $\uparrow$	Diversity LPIPS $\uparrow$	Quality Score $\uparrow$
Cosmos-Transfer1-7B [Vis]	<b>0.96</b>	0.16	0.49	<b>0.72</b>	0.19	5.94
Cosmos-Transfer1-7B [Edge]	0.77	<b>0.28</b>	0.53	0.71	0.28	5.48
Cosmos-Transfer1-7B [Depth]	0.71	0.14	0.49	0.70	<u>0.39</u>	6.51
Cosmos-Transfer1-7B [Seg]	0.66	0.11	0.75	0.68	<b>0.42</b>	6.30
Cosmos-Transfer1-7B Uniform Weights, no Vis	0.68	0.13	0.57	0.67	0.37	<u>8.02</u>
Cosmos-Transfer1-7B Uniform Weights, no Edge	0.81	0.10	0.53	0.66	0.31	7.68
Cosmos-Transfer1-7B Uniform Weights, no Depth	0.83	0.15	0.52	0.69	0.25	7.49
Cosmos-Transfer1-7B Uniform Weights, no Seg	0.84	0.15	<b>0.43</b>	0.70	0.23	7.83
Cosmos-Transfer1-7B Uniform Weights	<u>0.87</u>	<u>0.20</u>	0.47	<b>0.72</b>	0.22	<b>8.54</b>



**Figure 6: Diagram of spatiotemporal control weighting by different modalities** (Vis, Edge, Depth and Segmentation). The control weight maps are 0.0 in black pixel areas, and 0.5 in white areas. We note that while the caption broadly specifies a bicycle repair shop scene, the blue shirt with a white logo and the skin color of the man are maintained, due to these pixels being controlled by Vis and Edge. On the other hand, for the background controlled by Depth and Segmentation, the objects are positioned in the scene consistently but have their colors and textures randomized (e.g. red toolbox, yellow tripod, white repair stand). A new tool rack on the wall on the right is also added by the model.

and a weight of 0.5 to depth and segmentation respectively in the background. While a multitude of heuristics could be used, we choose this composition strategy to exemplify a situation in which the user wants to have close resemblance to the original videos for the salient objects in the scene but generate more diverse videos in the

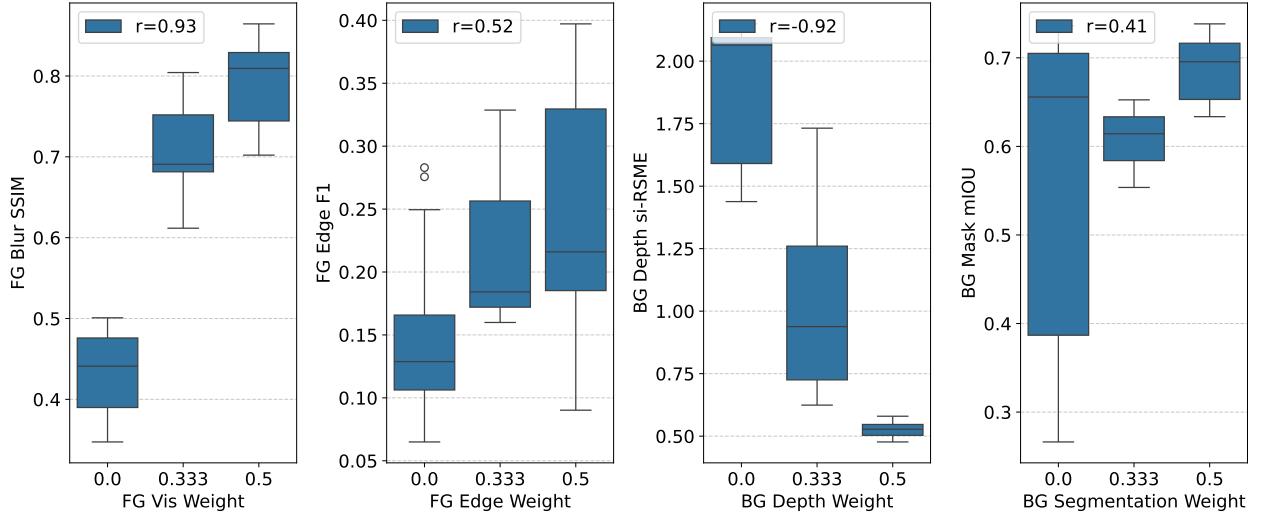


Figure 7: Correlations of modality weights on foreground (FG) region (for Vis and Edge) or background (BG) region (for Depth and Segmentation) with ground truth modality.

Table 2: Quantitative evaluation on TransferBench for Cosmos-Transfer1-7B with spatiotemporal weights derived from our *SalientObject* algorithm. The leftmost eight columns specify the weight design for the four modalities respectively. For each metric, “FG” denotes the result in that metric computed in the foreground region, and “BG” stands for background.

Control Weights								Vis Alignment	Edge Alignment	Depth Alignment	Segmentation Alignment	Diversity	Overall Quality					
FG				BG				Blur SSIM ↑ FG	Edge F1 ↑ BG	Depth si-RSME ↓ FG	Mask mIoU ↑ FG	Diversity LPIPS ↑ FG	Quality Score ↑ BG					
Vis	Edge	Depth	Seg	Vis	Edge	Depth	Seg											
0.5	0.5	0	0	0	0	0.5	0.5	0.81	0.71	0.27	0.14	0.37	0.52	0.77	0.68	0.01	0.33	8.29
0	0	0.5	0.5	0.5	0.5	0	0	0.68	0.93	0.17	0.25	0.38	0.40	0.77	0.75	0.12	0.03	8.08

background. Since vis and edge modalities are more constraining than depth and segmentation modalities, one can achieve this by using vis and edge modalities for the foreground, which leads to lower degree-of-freedom, and depth and segmentation modalities for the background, which have higher degree-of-freedom.

In addition to this qualitative example, we also conduct a series of quantitative experiments. The first experiment to demonstrate granular spatiotemporal control is shown in Fig. 7. Here, we condition the foreground with edge and vis, and background with depth and segmentation. We ablate the conditioning strength per modality (0, 0.333, 0.5) and show a strong correlation to the relevant alignment metric in the affected pixel regions. For example, by selectively increasing the foreground vis weight, we improve the foreground blur SSIM from 0.43 to 0.81, with a Pearson correlation coefficient of 0.93. Similarly, by selectively increasing the background depth weight, we improve the Depth si-RSME from 1.88 to 0.52, with a Pearson correlation coefficient of -0.92.

The second experiment we perform is shown in Tab. 2. Here, we use two different strategies to determine the spatiotemporal control map:

- We condition the foreground with edge and vis, and background with depth and segmentation.
- We condition the foreground with depth and segmentation, and background with edge and vis, essentially inverting foreground/background.

From Tab. 2, we observe that when swapping the edge and vis conditioning from foreground to background, there is a marked improvement in Blur SSIM and Edge F1 in the background and a degradation in the foreground. We also observe that vis and edge conditioning perform competitively for depth and segmentation

alignment, in agreement with Tab. 1.

Furthermore, we observe in Tab. 2 that when we swap the (foreground or background) region from vis and edge conditioning (low degree-of-freedom) to depth and segmentation conditioning (high degree-of-freedom), we observe a marked increase in diversity LPIPS score (FG 0.01 → 0.12, BG 0.03 → 0.33) and on-par visual quality, while also maintaining good depth and segmentation alignment (especially in reference to unconditioned baselines in Fig. 7). This matches our expectation from the qualitative example shown in Fig. 6.

### 5.3. Case Study for Robotics Sim2Real Data Generation

In robotics research, the availability of large-scale, high-quality data is critical, as supported by established scaling laws (Lin et al., 2025). While simulation provides a means to generate extensive datasets, the substantial domain gap between synthetic and real-world data presents significant challenges in transferring models trained in simulated environments to real-world applications (Muratore et al., 2022).

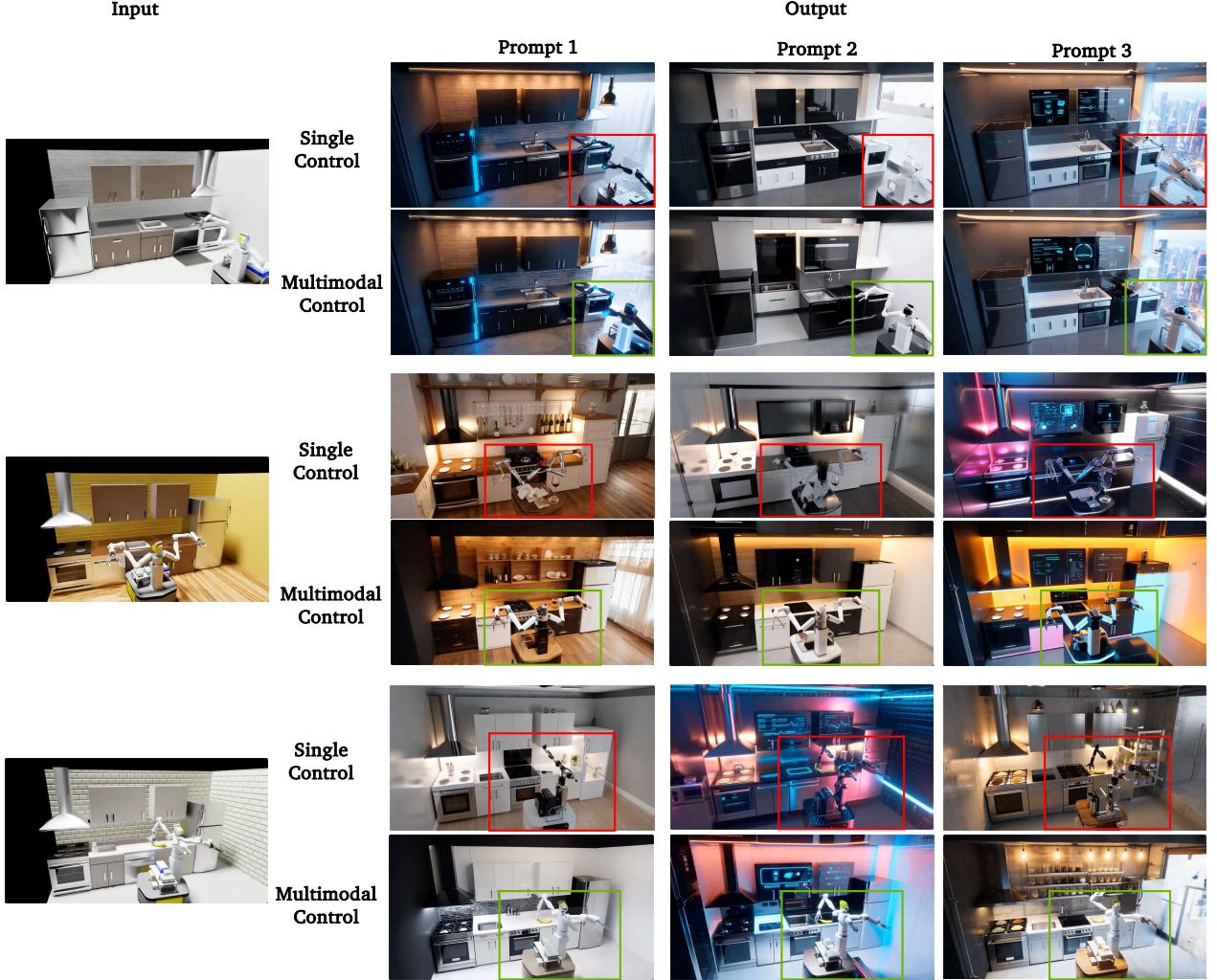
To evaluate Cosmos-Transfer1 for robotics data generation, we curated a small simulated dataset of 20 robot manipulation scenarios in a basic kitchen scene using NVIDIA Omniverse and Isaac Lab. The kitchen scenes were programmatically generated, and contain furniture with articulations for the robot to interact with. The tasks for the robot include opening and closing cabinets, and picking and placing common kitchen objects. The robot motions were computed by an automated task and motion planning algorithm (Garrett et al., 2020). For each of the scenarios, we prepared six different text prompts with descriptions of different illumination, scene details, and operating environments. In addition to RGB videos, we also output segmentation and depth maps for each scenario. In addition to the aforementioned four single-modal Cosmos-Transfer1-7B models, in order to fully preserve robot dynamics, we also employed the Cosmos-Transfer1-7B model with the spatiotemporal control map. Specifically, we first extracted the foreground robot masks using semantic segmentation results and then designed different modal constraint settings for both the foreground and background.

**Table 3: Quantitative evaluation of Cosmos-Transfer1 on robotics Sim2Real data generation task**, including single-control models and two multimodal control variants with different spatiotemporal control maps. Best results are in bold; second-best are underlined.

Model	Vis Alignment	Edge Alignment	Depth Alignment	Segmentation Alignment	FG Segmentation Alignment	Diversity	Overall Quality
	Blur SSIM ↑	Edge F1 ↑	Depth si-RMSE ↓	Mask mIoU ↑	FG Mask mIoU ↑	Diversity LPIPS ↑	Quality Score ↑
Cosmos-Transfer1-7B [Vis]	<b>0.95</b>	<u>0.19</u>	<b>0.82</b>	<b>0.65</b>	0.56	0.20	9.11
Cosmos-Transfer1-7B [Edge]	0.63	<b>0.40</b>	1.01	<u>0.63</u>	0.57	0.36	7.70
Cosmos-Transfer1-7B [Depth]	<u>0.66</u>	0.13	<u>0.84</u>	0.59	0.57	0.43	9.17
Cosmos-Transfer1-7B [Seg]	0.47	0.10	1.34	0.55	0.54	<b>0.60</b>	9.29
Cosmos-Transfer1-7B, Setting1	0.51	0.12	1.30	0.59	<u>0.61</u>	0.57	<u>9.57</u>
Cosmos-Transfer1-7B, Setting2	0.50	0.14	1.41	0.60	<b>0.63</b>	<u>0.58</u>	<b>10.42</b>

We propose two straightforward configurations for using Cosmos-Transfer1-7B. In Setting 1, our goal is to preserve both the shape and appearance of the foreground robot while modifying the background. To achieve this, we apply the edge and vis controls to the foreground (FG) while using the Segmentation control for the background (BG), assigning the following weights:  $w_{\text{Edge}}(\text{FG}) = 1$ ,  $w_{\text{vis}}(\text{FG}) = 1$ , and  $w_{\text{Seg}}(\text{BG}) = 1$ , with all other weights set to zero. In Setting 2, we aim to maintain only the shape of the robot while allowing variations in its appearance (e.g., color and texture). Therefore, we apply only the edge control to the foreground, while still using the segmentation control for the background, with the weights:  $w_{\text{Edge}}(\text{FG}) = 1$ , and  $w_{\text{Seg}}(\text{BG}) = 1$ , and all other weights set to zero.

Table 3 presents the quantitative evaluation of Cosmos-Transfer1 on robotics simulation data (120 videos, i.e.,  $20 \times 6$ ). The evaluation metrics include Blur SSIM, Edge F1, Mask mIoU, Diversity-LPIPS, and Quality Score. Additionally, we compute FG Mask mIoU for the foreground robot, excluding the background, to evaluate



**Figure 8: Example results of Cosmos-Transfer1 for robotic data generation.** The left column displays input videos generated by NVIDIA Isaac Lab, while the right three columns show results from Cosmos-Transfer1-7B with different condition modalities and spatiotemporal control maps. For each example, the top row (single) uses Segmentation as the condition modality with an overall constraint weight of 1. The bottom row combines Segmentation, Edge, and Vis as conditions, applying a spatiotemporal control map scheme. Specifically, a combination of Edge, Segmentation and Vis are used with a customized control weight on the foreground (robot region), while only segmentation with a control weight of 1 is applied to the background. These results demonstrate that Cosmos-Transfer1-7B with the spatiotemporal control map enhances the fidelity of the foreground robot.

the consistency between simulated and generated videos. As observed, single-modal Cosmos-Transfer1-7B models tend to yield higher scores for the corresponding metrics (e.g., Cosmos-Transfer1-7B [Vis] achieves the highest Blur SSIM) but result in a lower Quality Score than those from Cosmos-Transfer1-7B under the adaptive multimodal control settings. Cosmos-Transfer1-7B [Seg] produces higher Quality Scores and the highest Diversity-LPIPS but exhibits lower performance in Blur SSIM, Edge F1, and Mask mIoU, which suggests that Cosmos-Transfer1-7B [Seg] can generate diverse backgrounds but may introduce artifacts into the foreground robot. Overall, the two Cosmos-Transfer1-7B model settings are among the top three in Quality Score, Diversity-LPIPS, and FG Mask mIoU, while offering a more balanced performance across Blur SSIM, Edge F1, and Mask mIoU, which shows Cosmos-Transfer1-7B with spatiotemporal control map has better overall video quality, diversity, and the preservation of foreground robots. These statistical findings are further illustrated in the

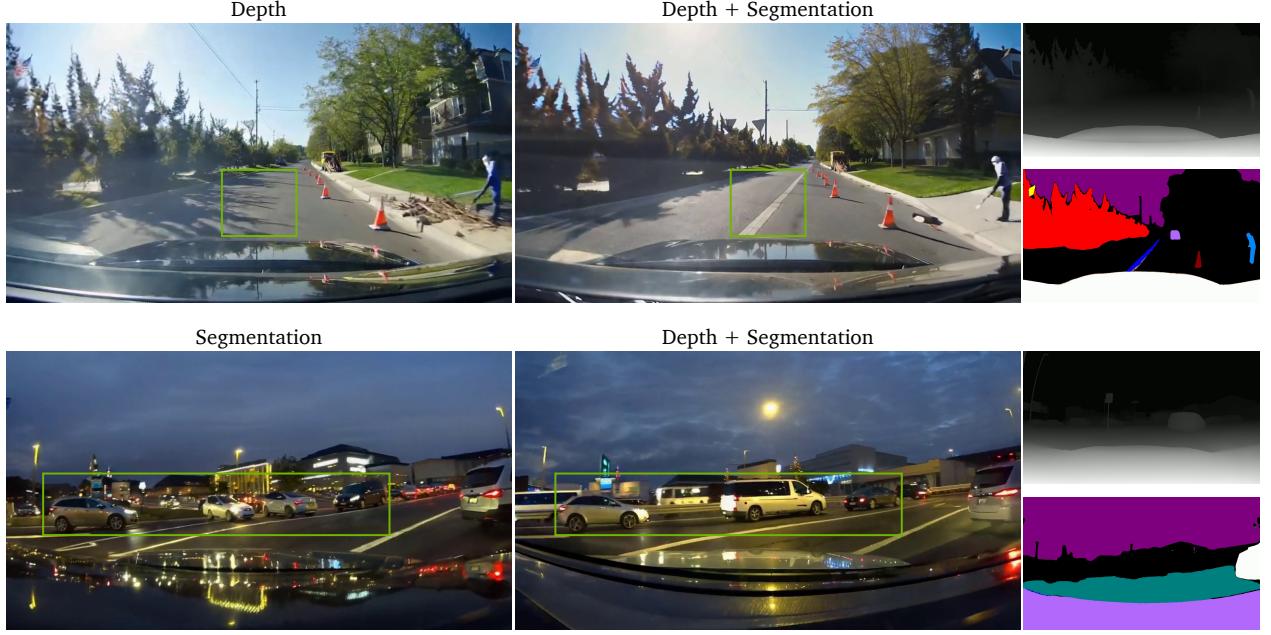


Figure 9: **Comparison of the generation results conditioned on depth and segmentation of Cosmos-Transfer1-7B.** In each example, the highlighted regions illustrate the enhancements achieved by incorporating multiple control signals over relying on a single one.

following figure.

Figure 8 shows three examples of using Cosmos-Transfer1-7B to augment simulated robotics data. The left column displays the input videos generated by NVIDIA Omniverse and Isaac Lab, while the right three columns show the outputs of Cosmos-Transfer1-7B with different condition modalities, spatiotemporal control maps, and text prompts. First, these results show Cosmos-Transfer1-7B significantly improves the photorealism and diversity of the robotics videos, by adding more scene details and complex shading and natural illumination. Moreover, as highlighted in the red and green rectangle regions, compared to using single-modal models, we found using the full Cosmos-Transfer1-7B model with spatiotemporal control maps on edge, segmentation and vis can better preserve the robot shape, reduce broken artifacts and result in overall better video quality. Please refer to the figure caption for more details.

As shown, Cosmos-Transfer1-7B offers a promising approach to mitigating the domain gap between simulation and real-world by enhancing the realism and diversity of synthetic data while preserving task-relevant properties. For instance, in robotics manipulation scenarios, Cosmos-Transfer1-7B can refine simulated videos by adjusting lighting, color, texture, and fine-grained scene details while maintaining physically plausible robot dynamics. Additionally, Cosmos-Transfer1-7B can enrich scene complexity by introducing novel background objects, thereby increasing the ecological validity of the data. By leveraging Cosmos-Transfer1-7B, researchers can improve the robustness and generalization of models trained in simulation, facilitating more effective deployment in real-world robotic environments.

#### 5.4. Case Study for Autonomous Driving Data Enrichment

Unlike robotics, the field of autonomous driving continuously accumulates vast amounts of real-world data collected by mass-produced vehicles and self-driving fleets. However, this data follows a highly long-tailed distribution, thus making it crucial to amplify the utilization of safety-critical corner cases. By leveraging Cosmos-Transfer1, autonomous vehicle developers can maximize the utility of real-world edge cases, ultimately leading to enriched and diversified driving data. Specifically, one can condition Cosmos-Transfer1 on the

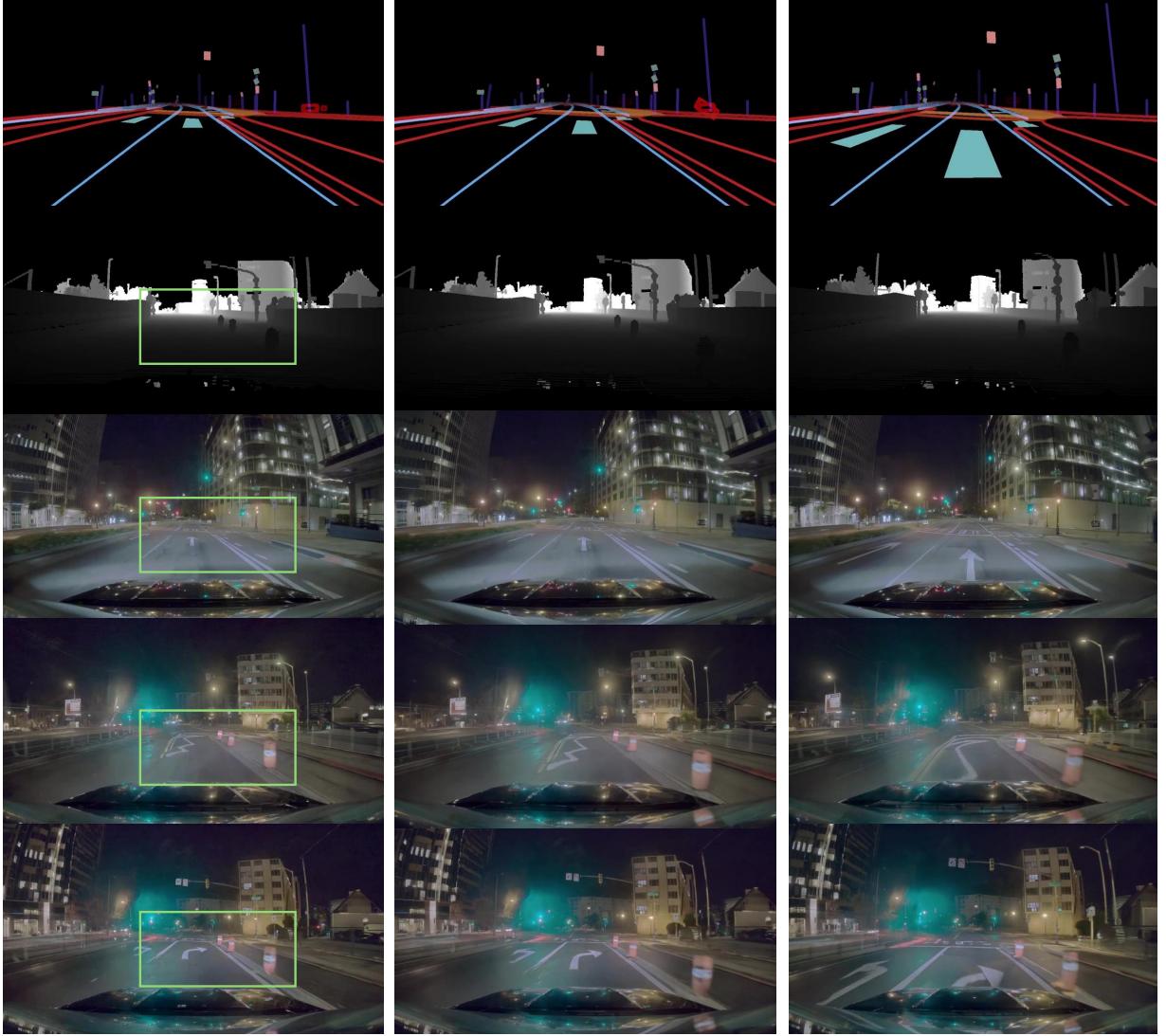


Figure 10: **Comparison of the generation results conditioned on HDMap and LiDAR of Cosmos-Transfer1-7B-Sample-AV.** The highlighted regions illustrate the enhancements achieved by incorporating multiple control signals compared to relying on a single one. **1st row:** HDMap condition. **2nd row:** LiDAR condition. **3rd row:** Video generated using only HDMap. **4th row:** Video generated using only LiDAR, where traffic cones are introduced by LiDAR, but lane markings are incorrect. **5th row:** Video generated using both HDMap and LiDAR, where the lane layout is improved and more detailed objects are synthesized.

interventions described in scenario formats to generate numerous useful visual variations for autonomous vehicle (AV) testing and training. This capability allows systematic exploration of challenging and rare scenarios, enhancing both testing coverage and training effectiveness.

Fig. 9 presents the comparison between generation results using a single control signal with Cosmos-Transfer1-7B—either Depth or Segmentation—and those obtained by combining both with Cosmos-Transfer1-7B. When integrating multiple control signals in the examples of the figure, we simply adopt a uniform spatiotemporal control map, setting  $w_{\text{depth}} = 0.5$  and  $w_{\text{segment}} = 0.5$ . In the first row, combining depth and segmentation restores the solid line in the middle of the road, which is however absent when using depth alone. This solid line is essential for creating a scenario where the ego vehicle must momentarily drive in the opposite lane to navigate around the safety zone delineated by traffic cones. In the second row, segmentation alone generates unrealistic vehicle headings within the same lane, possibly due to the lack of detailed geometric cues. By

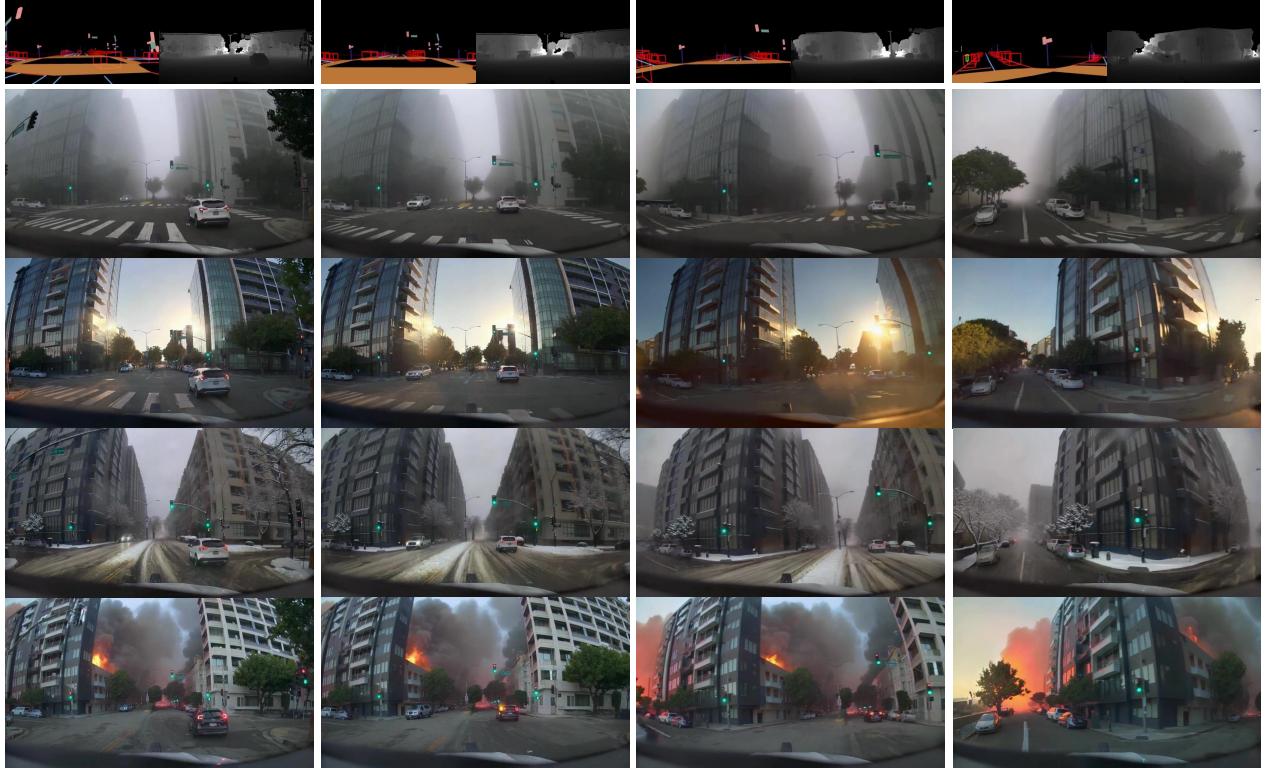


Figure 11: **1st row:** Control signals (left: HDMAP + 3DBbox, right: LiDAR) to Cosmos-Transfer1-7B-Sample-AV. **2nd-5th rows:** Video generated by different text prompts listed as following: *The scene unfolds on a foggy morning, with a thick layer of mist reducing visibility...;* *The scene is bathed in the warm, golden hues of the late afternoon sun, casting long shadows on the road...;* *The street is blanketed in heavy snowfall, with large snowflakes continuously falling, partially obscuring visibility...;* *The scene unfolds in a chaotic and intense environment as a fire engulfs the houses on either side of the street...*

integrating depth information, the fused depth and segmentation control signal produces more plausible vehicle orientations, improving the overall realism in the generated scene.

[Fig. 10](#) similarly compares a single control signal of HDMAP or LiDAR against their combined adaptive multimodal control signal based on Cosmos-Transfer1-7B-Sample-AV. In these experiments, we combine HDMAP and LiDAR control signals by setting uniform scalar values  $w_{\text{map}} = 0.3$  and  $w_{\text{lidar}} = 0.7$  for the spatiotemporal control map. In this example, the rightmost lane is blocked by traffic cones, and the ego vehicle attempts a right turn using the second right lane. In the fourth row, relying solely on LiDAR results in incorrect lane markings. In contrast, the fifth row demonstrates that fusing LiDAR with HDMAP significantly enhances the overall realism of the generated lane layout. Additionally, [Fig. 11](#) presents more examples illustrating various generation results by different text prompts.

We further assess the generation quality of a single control signal (HDMAP or LiDAR) and their combination using a diverse set of quantitative metrics in [Tab. 4](#). To assess adherence to the 3D object conditions, we utilize a 3D detection method built upon StreamPetr ([Wang et al., 2023](#)) and Hydra-MDP ([Li et al., 2024](#)). We compute the mean Average Precision (mAP) at an IoU threshold of 0.2 by comparing the predicted 3D bounding boxes on Cosmos-Transfer1-7B-Sample-AV-generated videos with their real-world counterparts. To measure adherence to the HDMAP, we employ the grounded SAM2 ([Liu et al., 2024; Ravi et al., 2024](#)) model for lane segmentation and compute the IoU between the predicted lane areas and the actual lane markings. Lastly, to quantitatively assess the 3D consistency of generated video with respect to the conditioned LiDAR data, we evaluate the photometric reprojection error following ([Zhou et al., 2017](#)). Specifically, we first subsample the

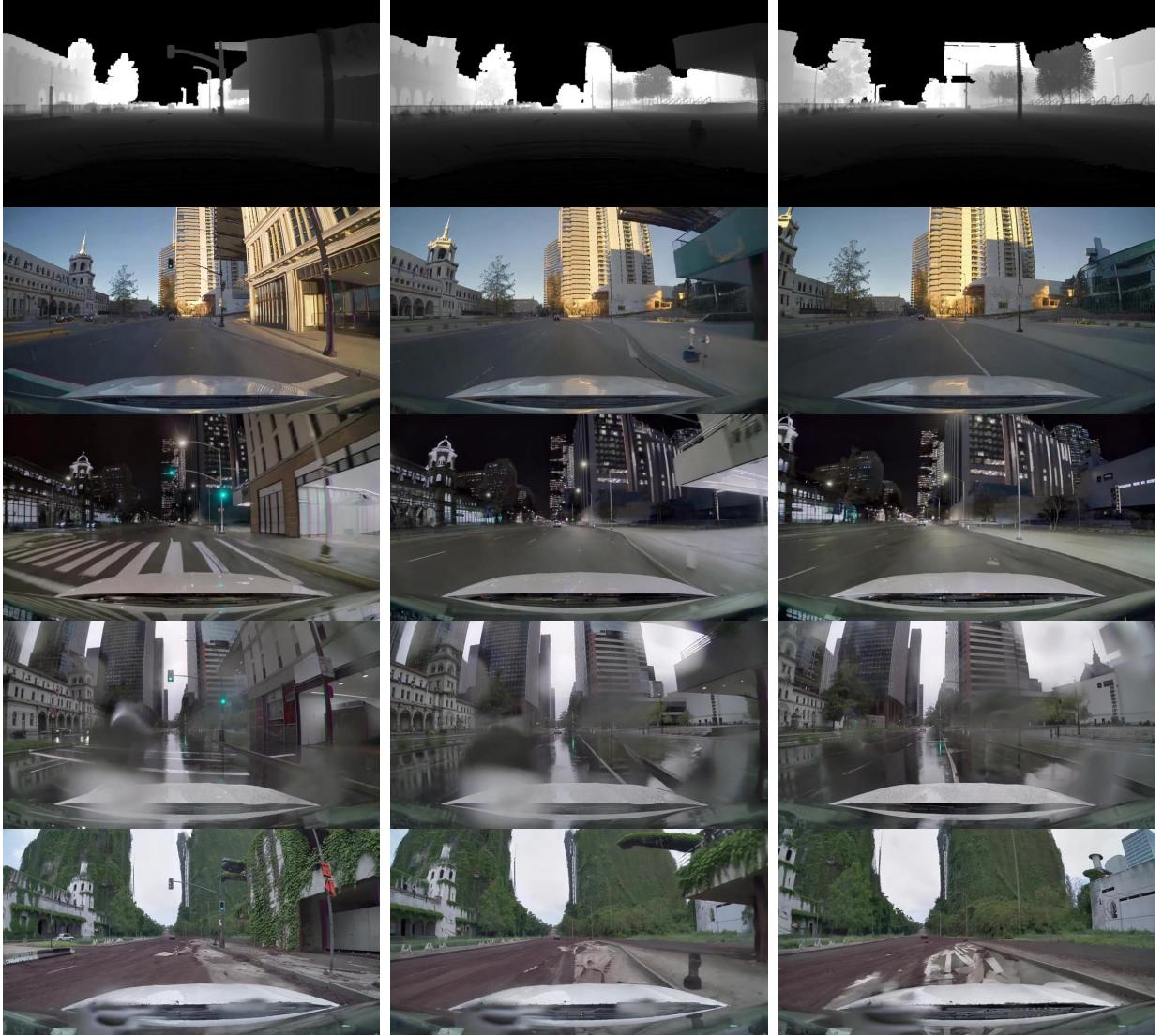


Figure 12: **1st row:** LiDAR simulated by NVIDIA Omniverse as the control signal to Cosmos-Transfer1-7B. **2nd-5th rows:** Videos generated by different text prompts listed as following: *The video showcases an urban driving scene during the golden hour...;*; *The video portrays a nighttime driving scene in an urban environment...;*; *The video captures an urban driving scene under heavy rainfall...;*; *The video depicts a thrilling driving scene in a jungle-style urban environment...*

generated video frames, retaining only those synchronized with the original LiDAR scans at 10 FPS. Using corresponding LiDAR-derived depths, known camera poses, and bounding box annotations, we warp each subsampled frame forward to the subsequent retained frame. We then compute the pixel-wise photometric discrepancy between each warped frame and the corresponding generated next frame using the L1 loss.

As shown in Tab. 4, Cosmos-Transfer1-7B-Sample-AV [LiDAR] achieves the best 3D-Bbox mAP scores, as it provides the most precise spatial conditioning signal. It also attains the lowest reprojection error, benefiting from LiDAR’s inherent 3D awareness. However, it performs the worst in lane mIoU due to the absence of explicit lane layout information. In contrast, Cosmos-Transfer1-7B-Sample-AV, which integrates both HDMap and LiDAR, achieves the highest lane mIoU while maintaining 3D-Bbox mAP, and its reprojection error score is comparable to Cosmos-Transfer1-7B-Sample-AV [LiDAR] and better than Cosmos-Transfer1-7B-Sample-AV [HDMap]. This highlights the benefits of integrating HDMap and LiDAR for improving scene generation,

**Table 4: Quantitative evaluation of Cosmos-Transfer1-7B-Sample-AV on the autonomous driving video generation task.** We compare the results of both single-control models and multimodal control variant over various metrics. Best results are in bold.

Method	3D-Bbox mAP ↑	Lane mIoU ↑	Reprojection Err. ↓
Cosmos-Transfer1-7B-Sample-AV [HDMap]	41.89	50.37	9.46
Cosmos-Transfer1-7B-Sample-AV [LiDAR]	<b>46.50</b>	48.19	<b>8.60</b>
Cosmos-Transfer1-7B-Sample-AV	44.66	<b>51.55</b>	8.67

particularly in enhancing lane accuracy—one of the key metrics critical to AV developers.

Finally, we demonstrate Cosmos-Transfer1-7B-Sample-AV for amplifying data variation in AV simulation. While previous examples rely on having a real-world counterpart of a scene layout, we can also connect Cosmos-Transfer1 with physically-based sensor simulation using the NVIDIA Omniverse Blueprint for AV simulation ([NVIDIA-Omniverse](#)). The blueprint is a reference workflow to create rich 3D worlds for AV training and testing. It contains APIs and services to model physics and behavior of dynamic objects in a scene and generate physically accurate and diverse sensor data.

In our experiment, we use Sensor RTX APIs in the blueprint to simulate LiDAR depth. As shown in [Fig. 12](#), we observe that, despite being trained on real LiDAR, the model generalizes well to synthetic LiDAR. As a bonus, the model effectively retains world knowledge from internet pre-training, enabling generation of scenes and effects not captured in our AV fine-tuning dataset. This demonstrates that Cosmos-Transfer1 can become an effective tool to enhance the visual diversity of simulated scenes.

## 6. Real-time Inference

In this section, we report an implementation of Cosmos-Transfer1-7B that achieves real-time inference performance by leveraging the new NVIDIA GB200 NVL72 system. **GB200 NVL72** binds together 36 Grace CPUs and 72 Blackwell GPUs, in an any-to-any NVLink network. This architecture is ideal for model parallelism techniques, such as Tensor and Context Parallelism. These techniques are used by many large-scale foundation models ([Grattafiori et al., 2024](#)), including World Foundation Models ([NVIDIA, 2025](#); [Parker-Holder et al., 2024](#)), in both training and inference scenarios. However, unlike Large Language Models (LLMs), which are often heavy in number of parameters and generate one token at a time, Cosmos-Transfer1-7B is relatively lightweight in number of parameters and generates tens of thousands of tokens in one shot.

Considering this use case, we devise a parallelism strategy for scaling Cosmos-Transfer1-7B in which we employ pure data parallelism in non-attention layers and head-parallelism in attention layers. Each B200 GPU packs up to 192GB of High-Bandwidth Memory (HBM), which can easily store an entire copy of the Cosmos-Transfer1-7B model. Therefore, to generate a 5-second 720p video, we can shard the entire 56K token sequence among GPUs. Under this setting, the only point of communication during diffusion is in attention, where we use the all-to-all collective so that each GPU operates on the entire 56K token sequence from a single attention head. This approach is favorable to the pipelined all-gathering of the key-value pair, since it does not require additional reduction steps and does not result in heavily memory-bandwidth-bound attention. Cosmos-Transfer1-7B has 32 attention heads, and uses classifier free guidance where the unconditional denoising is replaced with negative prompt-based denoising. We split the workload of the positive and negative conditionings among two groups of GPUs, as their denoising process is independent. As a result, we can distribute the attention workload among 64 GPUs so that each GPU performs attention over 56K query, key, and value tokens. Even with the largest query tile size used in Blackwell FMHA kernels ([Thakkar et al., 2023](#)), this still ensures full occupancy of the streaming multiprocessors (SMs) available in each B200 GPU.

We report generation times when using different numbers of GPUs in [Tab. 5](#). As shown, our parallelism strategy

**Table 5: Computation time for generating a 5-second video with Cosmos-Transfer1-7B under different parallelism settings.** End-to-end runtime dips below 5 seconds when scaled up to 64 B200 GPUs and reach real-time generation throughput.

Number of GPUs	1	4	8	16	32	64
Diffusion only	141.0 s	39.3 s	20.1 s	10.3 s	5.4 s	3.5 s
End-to-end	141.7 s	40.0 s	20.8 s	11.0 s	6.1 s	4.2 s

exhibits an approximately 40X speedup when going from 1 to 64 GPUs, when we only consider diffusion runtime, which is over 99% of the workload, and the only workload we parallelize across GPUs. We also achieve real-time generation throughput when using 64 GPUs, generating 5 seconds of video in only 4.2 seconds.

## 7. Related Work

**Visual Domain Transfer.** Numerous studies have explored the transfer of visual domains from abstract representations to photorealistic visualizations. A prominent line of research focuses on converting segmentation maps or sketches into high-fidelity images (Dundar et al., 2020; Fontanini et al., 2025; Huang et al., 2022; Lv et al., 2022, 2024; Park et al., 2019; Shi et al., 2022; Sushko et al., 2020; Wang et al., 2018, 2022, 2021; Xue et al., 2023). Beyond static images, several works extend this paradigm to video synthesis (Chung et al., 2023; Esser et al., 2023; Mallya et al., 2020; Wang et al., 2018, 2019; Zhuo et al., 2022, 2024), significantly enhancing the capability for dynamic and temporally coherent visual generation. These advancements have broad implications across various domains, including content creation, robotics, autonomous driving, virtual and augmented reality (VR/AR), and gaming.

**Spatial Control for Diffusion Models.** Diffusion models have demonstrated remarkable capabilities in text-to-image and text-to-video generation. To enhance their spatial controllability, various methods have been proposed, which can be broadly categorized into training-free approaches (Bansal et al., 2023; Chen et al., 2023; Xue et al., 2023) and techniques requiring additional training on top of pre-trained text-to-image models (Huang et al., 2023; Ju et al., 2023; Li et al., 2023; Liu et al., 2023, 2024; Lu et al., 2024; Mou et al., 2024; Qin et al., 2023; Ren et al., 2025; Zeng et al., 2023; Zhang et al., 2023; Zhao et al., 2024). Generally, the latter category achieves superior results. A notable work is ControlNet (Zhang et al., 2023), which introduced an additional encoder branch initialized from a pre-trained model, updating only this branch during training. Following its success, several extensions have been proposed (Ju et al., 2023; Qin et al., 2023; Sun et al., 2024; Zhao et al., 2024). More recently, Chen et al. (2024) extended the applicability of ControlNet from UNet-based architectures to transformers. Spatial control has also been explored in video generation. Lin et al. (2024) proposed to adapt a pre-trained ControlNet encoder from image to video. Jain et al. (2024) introduced a training-free approach leveraging masked attention modules to enable spatial control in video synthesis.

**Enhancing Simulation with Generative Models.** Developing and testing Physical AI systems in real-world environments pose significant risks, making simulation a crucial component for these tasks. Recent advancements in generative AI have greatly improved simulation by enhancing its realism, diversity, and utility. Early works leveraged generative models to refine simulator outputs. Rao et al. (2020) trained a CycleGAN to transform simulator-rendered images into more photorealistic counterparts while ensuring their utility by aligning the Q-values of the simulator and GAN-enhanced outputs. Ho et al. (2021) proposed RetinaGAN, an unsupervised GAN framework that enhances the realism of simulated scenes while preserving essential object features, demonstrating effectiveness in reinforcement learning across multiple real-world tasks. More recently, diffusion models have emerged as a powerful alternative to GANs for simulation enhancement. Zhao et al. (2024) utilized a diffusion-based model with ControlNet for sim-to-real transfer, converting simulation outputs (e.g. semantic maps) into photorealistic driving images. Their approach outperformed GAN-based methods, exhibiting fewer artifacts and better structural fidelity. Pronovost et al. (2023) introduced a scenario generation

pipeline based on latent diffusion, synthesizing complex traffic environments to enable scalable testing of autonomous agents under diverse and safety-critical driving conditions. Beyond static image synthesis, video generation models are increasingly being explored as learnable simulators for planning and control. NVIDIA (2025) presented a generative world model platform, providing a collection of open-source pre-trained models tailored for various physical AI tasks, further advancing simulation capabilities for real-world applications.

## 8. Conclusion

We introduced Cosmos-Transfer1, a diffusion-based conditional world model for multimodal controllable world generation. By introducing multimodal control branches to Cosmos-Predict1 (NVIDIA, 2025) with an adaptive weighting scheme, Cosmos-Transfer1 enables highly controllable world generation and improves generation quality. Extensive evaluations demonstrate that Cosmos-Transfer1 effectively preserves scene structure from the condition inputs while allowing fine-grained control, making it a powerful tool for bridging the synthetic-to-real domain gap in applications such as robotics Sim2Real and autonomous vehicle data enrichment. Additionally, our inference scaling strategy enables real-time throughput generation with an NVIDIA GB200 NVL72 rack, highlighting its practical feasibility. We open-sourced our code and models to help advance Physical AI research.

## A. Prompt Upsampler

One of the primary objectives of Cosmos-Transfer1 is to accommodate a diverse range of user requests, which can vary significantly in style, format, and content. During training, Cosmos-Transfer1 generates videos based on multimodal inputs and detailed descriptions. The detailed descriptions may differ substantially from user queries, leading to suboptimal performance. To mitigate this domain shift, we train a prompt upsampler to align diverse input prompts with the training prompt distribution of Cosmos-Transfer1.

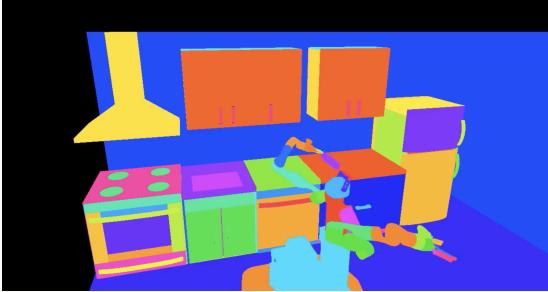
Following NVIDIA (2025), we aim to develop prompt upsamplers that satisfy the following three key criteria

- **Fidelity.** The upsampled prompt should remain faithful to the original short prompt and its corresponding conditioned video, avoiding any semantic conflicts.
- **Completeness.** The expanded prompt should fully retain and elaborate upon all elements present in the original prompt.
- **In-Distribution Consistency.** The structure and length of the upsampled prompt should align with the distribution of long prompts in the training dataset, ensuring optimal performance.

Specifically, we finetune one Pixtral-12B (Agrawal et al., 2024) to take in both the condition video and a user prompt, transforming it into a more detailed prompt with consistent structure. The input condition video can be of different modalities (e.g. segmentation mask, depth). To curate high-quality training data for this task, we begin with existing training prompts and utilize Gemma-2-9B-it (Gemma Team, 2024) to generate diverse short prompts that preserve fidelity while exhibiting varied characteristics. We curate a paired dataset of 1M videos for each modality and jointly train on all modality videos for one epoch with FSDP2 (Zhao et al., 2023).

For example, the prompt upsampler can understand and recognize key elements from segmentation maps and transform input text with rich information:

**EXAMPLE OUTPUT FROM UPSAMPLER THAT TAKES IN SHORT PROMPT AND A SEGMENTATION MAP, AND TRANSFORM INTO LONGER PROMPTS WITH DETAILS.**

<p>A robot in the kitchen picks up a bottle from the floor and puts it on a table.</p> 	<p><b>OUTPUT</b> The video features a kitchen with wooden cabinets and a granite countertop. A robot with a white body, black joints, and a red light on its head is seen performing tasks. It moves its arms and legs to pick up a white bottle with a red label from the floor and place it on the countertop. The robot then moves to a dining area with a wooden table and chairs, where it picks up a white chair and places it back in its original position.</p>
--	---

In a similar spirit, for depth input, the prompt upsampler also develops a holistic understanding of the scene and can recognize key surrounding objects:

EXAMPLE OUTPUT FROM UPSAMPLER THAT TAKES IN SHORT PROMPT AND DEPTH, AND TRANSFORM INTO LONGER PROMPTS WITH DETAILS.

A car is driving on a road.



OUTPUT

The video is taken from the perspective of a car's dashboard, showing the view of the road ahead. It is daytime with clear skies. The road is surrounded by greenery, with trees and shrubs visible on both sides. The car is driving on a road with multiple lanes, and there are other vehicles visible in the video, including a motorcycle. The lighting suggests it is daytime, and the weather appears to be clear.

## B. Contributors and Acknowledgements

### B.1. Core Contributors

Maciej Bala, Tiffany Cai, Francesco Ferroni, Sanja Fidler, Dieter Fox, Yunhao Ge, Jinwei Gu, Ali Hassani, Pooya Jannaty, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Qianli Ma, Hanzi Mao, Fabio Ramos, Xuanchi Ren, Tianchang Shen, Shitao Tang, Ting-Chun Wang, Jiashu Xu, Xiaodong Yang, Xiaohui Zeng, Yu Zeng

**Contributions:** **MYL** initiated the adaptive Multimodal Control design. **TCW** and **XZ** trained the individual ControlNets of Cosmos-Transfer1-7B. **TCW**, **TC**, **PJ** implemented the Multimodal Control inference. **ST**, **YG**, **QM**, **HM** contributed to the Cosmos-Transfer1-7B training data curation. **FR**, **DF**, **JG** and **YG** curated robotics simulation data. **YL**, **XR**, **TS** curated the RDS-HQ dataset. **QM**, **HM**, **MYL** designed the evaluation framework. **FF** and **QM** built the evaluation benchmark and framework. **QM**, **FF**, **JG**, **YG**, **XY**, **HM** conducted evaluations for Cosmos-Transfer1-7B. **XR**, **TS**, **HL** trained the individual ControlNets of Cosmos-Transfer1-7B-Sample-AV and conducted corresponding experiments. **JX** and **YG** built the prompt upsampler. **AH** and **MB** designed and implemented real-time inference with the NVIDIA GB200 NVL72 system. **SF** and **HL** supervised the Cosmos-Transfer1-7B-Sample-AV research, data curation, model training and evaluation. **HM** supervised the Cosmos-Transfer1-7B research, data curation, model training and evaluation. All core authors contributed to paper writing. **HM** and **MYL** organized paper writing. **MYL** supervised the project.

### B.2. Contributors

Hassan Abu Alhaija, Jose Alvarez, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Michael Isaev, Shiyi Lan, Tobias Lasser, Alice Luo, Xinglong Sun, Jay Wu, Kevin Xie, Stella Xu, Yuchong Ye

### B.3. Acknowledgments

Joshua Bapst, Prithvijit Chattopadhyay, TJ Galda, Mingfei Guo, Madison Huang, Hai Loc Lu

## References

- [1] OpenAI: Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>. 9
- [2] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems, 2025. 7
- [3] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Moncault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 21
- [4] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *CVPR*, 2023. 19
- [5] John Canny. A computational approach to edge detection. *TPAMI*, 1986. 4
- [6] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart- $\{\delta\}$ : Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. 2, 19
- [7] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 19
- [8] Chaeyeon Chung, Yeojeong Park, Seunghwan Choi, Munkhsoyol Ganbat, and Jaegul Choo. Shortcut-v2v: Compression framework for video-to-video translation based on temporal redundancy reduction. In *ICCV*, 2023. 19
- [9] Aysegul Dundar, Ming-Yu Liu, Zhiding Yu, Ting-Chun Wang, John Zedlewski, , and Jan Kautz. Domain stylization: A fast covariance matching framework towards domain adaptation. 2020. 19
- [10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 2014. 8
- [11] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023. 19
- [12] Tomaso Fontanini, Claudio Ferrari, Giuseppe Lisanti, Massimo Bertozzi, and Andrea Prati. Semantic image synthesis via class-adaptive cross-attention. *IEEE Access*, 2025. 19
- [13] Caelan Reed Garrett, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning. In *Proceedings of the international conference on automated planning and scheduling*, volume 30, pages 440–448, 2020. 12
- [14] Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 21
- [15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 18
- [16] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives, 2024. URL <https://arxiv.org/abs/2311.18259>. 7

- [17] Daniel Ho, Kanishka Rao, Zhuo Xu, Eric Jang, Mohi Khansari, and Yunfei Bai. Retinagan: An object-aware approach to sim-to-real transfer. In *ICRA*, 2021. 19
- [18] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 19
- [19] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts GANs. In *ECCV*, 2022. 19
- [20] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *CVPR*, 2024. 19
- [21] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *ICCV*, 2023. 19
- [22] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 19
- [23] Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024. 16
- [24] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation, 2025. URL <https://arxiv.org/abs/2410.18647>. 12
- [25] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024. 19
- [26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 4, 16
- [27] 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. *arXiv preprint arXiv:2310.08579*, 2023. 19
- [28] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. In *CVPR*, 2024. 19
- [29] Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, and Jiahui Huang. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models, 2024. URL <https://arxiv.org/abs/2412.03934>. 19
- [30] Zhengyao Lv, Xiaoming Li, Zhenxing Niu, Bing Cao, and Wangmeng Zuo. Semantic-shape adaptive feature modulation for semantic image synthesis. In *CVPR*, 2022. 19
- [31] Zhengyao Lv, Yuxiang Wei, Wangmeng Zuo, and Kwan-Yee K Wong. Place: Adaptive layout-semantic fusion for semantic image synthesis. In *CVPR*, 2024. 19
- [32] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. In *ECCV*, 2020. 19
- [33] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 19

- [34] Fabio Muratore, Fabio Ramos, Greg Turk, Wenhao Yu, Michael Gienger, and Jan Peters. Robot learning from randomized simulations: A review. *Frontiers in Robotics and AI*, 2022. 12
- [35] NVIDIA. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1, 3, 4, 18, 20, 21
- [36] NVIDIA-Omniverse. Omniverse blueprint for av simulation. [cases/autonomous-vehicle-simulation/?deeplink=content-tab--3](#). 18
- [37] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 19
- [38] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, et al. Genie 2: A large-scale foundation world model. 2024. URL <https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/>. 18
- [39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1
- [40] Ethan Pronovost, Meghana Reddy Ganesina, Noureldin Hendy, Zeyu Wang, Andres Morales, Kai Wang, and Nick Roy. Scenario diffusion: Controllable driving scenario generation with diffusion. *NeurIPS*, 2023. 19
- [41] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 19
- [42] Kanishka Rao, Chris Harris, Alex Irpan, Sergey Levine, Julian Ibarz, and Mohi Khansari. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11157–11166, 2020. 19
- [43] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4, 8, 16
- [44] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, 2025. 19
- [45] Yupeng Shi, Xiao Liu, Yuxiang Wei, Zhongqin Wu, and Wangmeng Zuo. Retrieval-based spatially adaptive normalization for semantic image synthesis. In *CVPR*, 2022. 19
- [46] Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: create your artwork with versatile control on text-to-image generation. In *ECCV*, 2024. 19
- [47] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020. 19
- [48] Vijay Thakkar, Pradeep Ramani, Cris Cecka, Aniket Shivam, Honghao Lu, Ethan Yan, Jack Kosaian, Mark Hoemmen, Haicheng Wu, Andrew Kerr, et al. Cutlass, 2023. URL <https://github.com/NVIDIA/cutlass>. 18
- [49] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *ICCV*, 1998. 4
- [50] C.J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, 1979. 8

- [51] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *arXiv preprint arXiv:2303.11926*, 2023. 16
- [52] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018. 19
- [53] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 19
- [54] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *NeurIPS*, 2019. 19
- [55] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 19
- [56] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 6
- [57] Yi Wang, Lu Qi, Ying-Cong Chen, Xiangyu Zhang, and Jiaya Jia. Image synthesis via semantic composition. In *ICCV*, 2021. 19
- [58] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 8
- [59] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023. 8
- [60] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *CVPR*, 2023. 19
- [61] 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 *CVPR*, 2024. 7
- [62] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 4, 8
- [63] Yu Zeng, Zhe Lin, Jianming Zhang, Qing Liu, John Collomosse, Jason Kuen, and Vishal M Patel. Scenecomposer: Any-level semantic image synthesis. In *CVPR*, 2023. 19
- [64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1, 19
- [65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 8
- [66] Haonan Zhao, Yiting Wang, Thomas Bashford-Rogers, Valentina Donzella, and Kurt Debattista. Exploring generative ai for sim2real in driving data synthesis. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024. 19
- [67] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2024. 19
- [68] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023. 21

- [69] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. [16](#)
- [70] Long Zhuo, Guangcong Wang, Shikai Li, Wayne Wu, and Ziwei Liu. Fast-vid2vid: Spatial-temporal compression for video-to-video synthesis. In *ECCV*, 2022. [19](#)
- [71] Long Zhuo, Guangcong Wang, Shikai Li, Wayne Wu, and Ziwei Liu. Fast-vid2vid++: Spatial-temporal distillation for real-time video-to-video synthesis. *TPAMI*, 2024. [19](#)