Unpaired Geometry-Guided Sim2Real Translation for Autonomous Driving

Xinzhuo Chen[1,*], Shijie Wang[3,*], Chao Gao[1,4], Jinguang Gu[1,2,†], and Gongjin Lan[3,†]

[1] School of Computer Science and Technology, Wuhan University of Science and Technology, China
[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, China
[3] Department of Computer Science and Engineering, Southern University of Science and Technology, China
[4] Vrije Universiteit Amsterdam, The Netherlands
{chenxinzhuo,simon}@wust.edu.cn
wangsj2024@mail.sustech.edu.cn

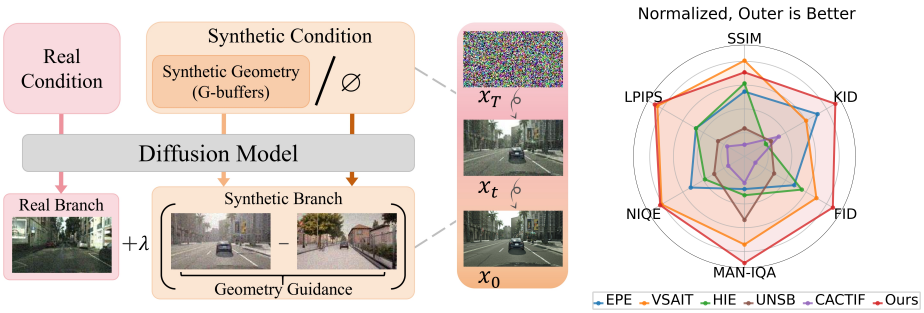
Abstract. Simulation is indispensable for autonomous driving, yet the pronounced visual Sim2Real gap severely limits the cross-domain generalization of perception models. To bridge this gap, we propose Cross-Domain Control Transfer (CDCT), a diffusion-based Sim2Real framework for translating synthetic images into realistic counterparts without paired supervision. CDCT features a novel cross-domain score composition mechanism that injects domain-agnostic geometry guidance, derived from synthetic-domain score differences, into a real-world appearance prior. This ensures strict geometric and semantic consistency without paired real-world spatial conditions. However, naively incorporating multiple geometric conditions often incurs a prohibitive parameter overhead. To address this, our Lightweight Multi-Condition Adapter (LMCA) module processes diverse rendering buffers simultaneously, eliminating the computational redundancy of standard multi-branch architectures. Extensive experiments on the CARLA-to-Cityscapes benchmark demonstrate that CDCT yields substantial zero-shot performance gains on downstream perception tasks. Crucially, joint training with our generated data surpasses the performance of real-world-only training, showcasing its immense potential for autonomous driving.

Keywords: Sim2Real Gap · Diffusion Model · Autonomous Driving

1 Introduction

Simulation provides a critical, controllable platform for training autonomous driving perception models. Crucially, it enables the evaluation of rare corner cases and facilitates large-scale data generation with precise, cost-free annotations that are prohibitively expensive to acquire in the real world.

* Equal contribution, †Co-corresponding authors.



(a) Overview of Cross-Domain Control Transfer **(b)** Performance Comparison

Fig. 1: (a) Cross-Domain Control Transfer bridges the Sim2Real gap by injecting geometry guidance from the Synthetic Branch into the real-world appearance prior of the Real Branch. This enables the generation of high-fidelity images that strictly preserve simulated semantic layouts without paired supervision. (b) Performance comparison with Sim2Real methods across multiple metrics.

Despite rapid advances in computer graphics and rendering technologies, a substantial Sim2Real gap [16, 29, 44] fundamentally limits the zero-shot deployment of perception models trained purely on synthetic data. Prior analyses categorize this discrepancy into two complementary aspects: the content gap [18] and the appearance gap [29]. The content gap refers to differences in scene composition, object distribution, and behavioral patterns between simulated and real-world data. In contrast, the appearance gap stems from discrepancies in low-level image formation factors, such as texture statistics, illumination distributions, and material reflectance properties. Modern deep perception models are profoundly sensitive to these low-level statistics, causing severe performance degradation during real-world deployment. Consequently, bridging this appearance gap remains a pervasive bottleneck.

Recently, generative models, particularly diffusion models, have made image-to-image translation a promising approach for narrowing the appearance gap [50], owing to their superior capacity to represent complex textures and illumination statistics in real-world scenes. However, two critical challenges arise when extending generative models to Sim2Real translation under autonomous driving scenarios.

First, the dilemma of unpaired conditional guidance. To ensure strict geometry consistency, conventional Sim2Real translation paradigms heavily rely on paired simulated-real data for pixel-aligned supervision, which is practically impossible to satisfy in dynamic, large-scale driving environments. While perfect geometry conditions like G-buffers are abundant in simulators, the core difficulty lies in establishing a mapping that utilizes precise simulation geometry to guide real-world texture synthesis without access to ground-truth real-world conditions. **Second, the computational redundancy in multi-condition control.** Complex driving scenes require diverse geometry constraints; a single modality is often insufficient. While multi-condition frameworks exist, naively

stacking control branches leads to prohibitive computational costs and memory usage.

In response to these challenges, we propose Cross-Domain Control Transfer (CDCT), a multi-conditional geometry-aware Sim2Real diffusion framework as shown in Fig. 1. Our contributions are:

1. **Cross-domain score composition for unpaired geometry control.** We derive a cross-domain score composition function that estimates a domain-agnostic geometry guidance term from synthetic-domain score differences and injects it into a real-domain appearance prior during sampling, enabling geometry-consistent Sim2Real translation without paired real conditions.
2. **Parameter-efficient multi-condition control.** We introduce the Lightweight Multi-Condition Adapter (LMCA), trained on synthetic G-buffers with a frozen backbone, supporting dense multi-condition guidance at high resolution while avoiding the redundancy of stacking multiple control branches. Specifically, our LMCA introduces only **2.8%** additional parameters per condition, in stark contrast to the massive overhead of standard control paradigms.
3. **Superior data utility for downstream perception tasks.** Extensive experiments on the CARLA-to-Cityscapes benchmark demonstrate the superiority of CDCT over recent Sim2Real baselines. By narrowing the appearance gap, models trained solely on our generated data achieve remarkable zero-shot performance gains across multiple perception tasks, including semantic segmentation, object detection, and monocular depth estimation (+33.7 **mIoU** in semantic segmentation, +20.3 $\mathbf{mAP}_{50}$ in object detection, and reducing **AbsRel** from 0.56 to 0.16 in depth estimation compared to simulation-only training). Crucially, augmenting real-world datasets with our translated data further surpasses the performance of models trained exclusively on real data, validating the practical value of CDCT for scalable autonomous driving perception.

2 Related Work

Unpaired image-to-image translation for Sim2Real. Sim2Real translation is commonly formulated as an unpaired image-to-image translation problem, since paired synthetic–real images with identical geometry are difficult to obtain. Generative image-to-image translation models have therefore been widely adopted to bridge the Sim2Real appearance gap [5, 23, 32, 46, 50]. Early GAN-based [10] unpaired methods, such as CycleGAN [52] and CUT [30], eliminate the requirement for strictly paired data. However, lacking explicit spatial constraints, these methods tend to inadvertently alter object layouts.

To improve structural consistency, subsequent studies enforce task-level semantic consistency [14] or incorporate intermediate G-buffers [1,32] for geometric guidance. Beyond external conditioning, other works have explored specialized architectures and generative frameworks to prevent semantic distortion. Specifically, Hierarchy Flow [9] utilizes reversible normalizing flows [20] for lossless

feature extraction and reconstruction, VSAIT [40] leverages vector symbolic architectures to explicitly preserve structural bindings, and UNSB [19] formulates the translation as a Neural Schrödinger Bridge regularized by contrastive learning to maintain structural integrity across the iterative generation process.

Although preserving layout consistency, these methods struggle with unstable optimization and limited capacity, falling short in synthesizing complex, photo-realistic appearances.

Recently, diffusion models [12,33] have achieved state-of-the-art performance in generative modeling. Nevertheless, few studies [5, 50] have explored the application of diffusion models to Sim2Real translation. Furthermore, existing unpaired diffusion frameworks such as DDIBs [38] and CycleDiff [53] primarily rely on probability flow ODEs or cyclic constraints. Lacking explicit geometric guidance, these methods frequently introduce semantic distortions and struggle to strictly preserve the complex spatial layouts required for autonomous driving.

Controllable Image Generation with Diffusion. To guide diffusion models beyond text prompts, foundational works like ControlNet [47] and others [22, 27, 28, 36, 37, 51] directly align conditional inputs (e.g., depth, semantic maps) with the generated output and enforce pixel-level alignment. The architectural shift to Diffusion Transformers (DiT) [31], which surpass conventional UNet-based design [34] in scalability and representation capacity, has enabled significant progress. Recent advancements like OminiControl [39] propose parameter-efficient architectures, and frameworks like UniCombine [41] and EasyControl [49] enable multi-conditional generation. However, they are primarily designed for open-domain artistic synthesis and fail to preserve the precise scene geometry required in autonomous driving Sim2Real translation.

3 Method

3.1 Preliminaries

Our framework builds upon the latest generation of diffusion models, specifically those utilizing the Multi-Modal Diffusion Transformer (MM-DiT) [8] architecture and the Rectified Flow [24, 25] training strategy such as Stable Diffusion 3 [8] and Flux.1 [21]. Instead of the single-branch architecture [33] where the text prompt is injected into the denoising branch via cross-attention, MM-DiT uses two independent transformers to construct the text branch and the image denoising branch.

Within each MM-DiT block, noise image tokens X and text tokens T are projected into their respective queries Q, keys K, and values V representations. It enables the computation of attention between all tokens by Multi-Modal Attention (MMA):

$$\mathrm{MMA}[X;T] = \mathrm{softmax}\left(\frac{1}{\sqrt{d}} Q_{\mathrm{x\&t}} K_{\mathrm{x\&t}}^{\top}\right) V_{\mathrm{x\&t}} \tag{1}$$

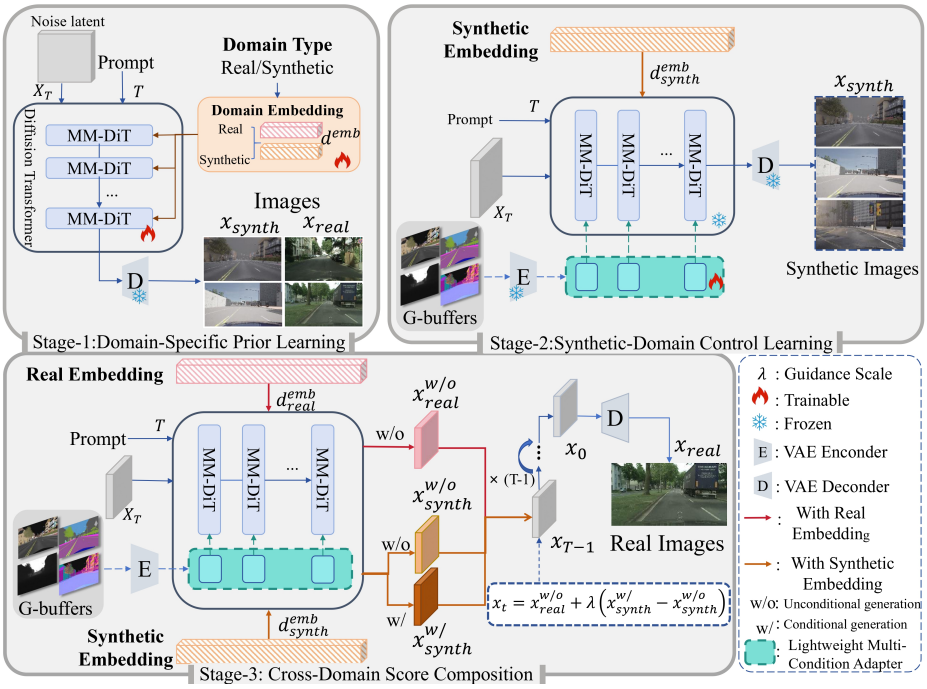


Fig. 2: Overview of CDCT framework. Stage-1 learns domain-specific appearance priors via pretrained diffusion fine-tuning on unpaired real and synthetic data. Stage-2 trains a control module on synthetic G-buffers to capture geometry-conditioned score modulation while freezing the backbone. Stage-3 transfers geometric control to the real-domain prior through score composition, enabling geometry-consistent Sim2Real synthesis without paired supervision.

where $[X;T]$ denotes the concatenation of image and text tokens. This formulation enables bidirectional attention.

3.2 Cross-Domain Control Transfer

Problem Formulation. Let $x \in \mathcal{X}$ be an image and $g \in \mathcal{G}$ be structured geometry conditions extracted from simulation. We consider two domains indicated by $d \in \{d_{\text{synth}}, d_{\text{real}}\}$, where d_{synth} denotes synthetic renderings and d_{real} denotes real images. Our training set is *unpaired*: we have synthetic images with geometry annotations (x_{synth}, g) and real images without g; no synthetic–real correspondences are assumed. In the framework of score-based generative models such as diffusion models, the sampling distribution is characterized by its score function [35] and conditional generation can be achieved by composing score estimates from different conditioning variables [13]. The objective is to sample from $p(x \mid d_{\text{real}}, g)$, i.e., produce real-looking images while preserving the geometry specified by g.

The fundamental challenge lies in the domain mismatch: geometry supervision is only available in the synthetic domain, but the target distribution is real-domain conditional generation. As shown in Fig. 2, our framework tackles this through three sequential stages: (1) learn domain-specific appearance priors in a unified diffusion backbone, (2) learn geometry controllability *only* where supervision exists (synthetic), and (3) transfer the learned geometric control to real-domain sampling by composing scores during inference.

Domain-Specific Prior Learning. In Stage-1, we fine-tune a pretrained diffusion backbone on the unpaired real and synthetic datasets. We inject the discrete domain indicator $d \in \{d_{\text{real}}, d_{\text{synth}}\}$ through a lightweight learnable embedder (a linear projection) that maps d to a continuous vector d^{emb}, which is broadcast to all MM-DiT blocks. Since we utilize the SD3 backbone, the model is optimized using the Rectified Flow velocity prediction objective, learning $v_\theta(z_t, t, d)$, where z_t is the noise latent at timestep t.

This stage aligns a unified backbone to model both geometry-marginalized distributions, $p(x \mid d_{\text{real}})$ and $p(x \mid d_{\text{synth}})$. By switching d, the model can generate images with either real or synthetic appearances.

However, this generation is purely appearance-driven and geometrically unconstrained, which leads to the generated scene structure remaining stochastic. To achieve controllable generation, we must introduce explicit geometry guidance, which is the focus of the next stage.

Synthetic-Domain Control Learning. To enforce geometry constraints, Stage-2 leverages the paired synthetic data (x_{synth}, g). We freeze the backbone trained in Stage-1 and introduce a control branch. This branch encodes the structured G-buffer g and injects the conditional features into the backbone. During this stage, the domain embedding is fixed to d_{synth}.

Freezing the backbone preserves the domain-specific appearance priors learned in Stage-1 and forces the control module to capture geometry-conditioned modulation. The model now effectively samples from $p(x \mid d_{\text{synth}}, g)$.

Although we achieve controllable generation in the synthetic domain, this control does not trivially generalize. If we directly switch d to d_{real} and apply the same control module during inference, the generation typically fails. The control module was optimized solely under synthetic statistics and suffers from severe domain shift. This necessitates a principled approach to transfer the synthetic control signal without requiring the control module to generalize across domains.

Cross-Domain Score Composition. Our goal is to sample from the unseen distribution $p(x \mid d_{\text{real}}, g)$. We observe that Stage-1 provides a reliable real-domain appearance prior d_{real}, while Stage-2 provides accurate geometric modulation solely under the synthetic domain d_{synth}. Rather than forcing the control module to generalize across domains, we achieve transfer directly in the score space.

We assume: given the image x, its domain indicator d and underlying geometry g are conditionally independent. That is,

$$p(d, g \mid x) = p(d \mid x)\, p(g \mid x) \tag{2}$$

This assumption is reasonable in our setting because geometry is deterministically derived from scene structure and is independent of the domain indicator. Under this assumption, applying Bayes' rule yields the conditional score decomposition:

$$\nabla_x \log p(x \mid d, g) = \nabla_x \log p(x) + \nabla_x \log p(d \mid x) + \nabla_x \log p(g \mid x) \tag{3}$$

For single-domain conditioning, we naturally have:

$$\nabla_x \log p(x \mid d) = \nabla_x \log p(x) + \nabla_x \log p(d \mid x) \tag{4}$$

Combining the above equations, we can isolate the geometry term:

$$\nabla_x \log p(x \mid d, g) = \nabla_x \log p(x \mid d) + \nabla_x \log p(g \mid x) \tag{5}$$

Crucially, the geometry score $\nabla_x \log p(g \mid x)$ is independent of d. Therefore, we can reliably estimate it using the conditional and unconditional predictions from the synthetic domain:

$$\nabla_x \log p(g \mid x) \approx \nabla_x \log p(x \mid d_{\text{synth}}, g) - \nabla_x \log p(x \mid d_{\text{synth}}) \tag{6}$$

Substituting this back into Eq. (4) for the real domain ($d = d_{\text{real}}$), we derive our cross-domain inference formulation:

$$\begin{aligned} \nabla_x \log p(x \mid d_{\text{real}}, g) = {} & \nabla_x \log p(x \mid d_{\text{real}}) \\ & + \lambda \underbrace{\Big(\nabla_x \log p(x \mid d_{\text{synth}}, g) - \nabla_x \log p(x \mid d_{\text{synth}})\Big)}_{\text{Geometry Guidance}} \end{aligned} \tag{7}$$

where λ controls the strength of geometry guidance.

Implementation. In practice, for flow-based models like SD3, this score composition is equivalently implemented using velocity predictions. We run three forward passes at each timestep: (i) an unconditional real branch $v_\theta(x_t, t, d_{\text{real}})$, (ii) an unconditional synthetic branch $v_\theta(x_t, t, d_{\text{synth}})$, and (iii) a conditional synthetic branch with control module $v_\theta(x_t, t, d_{\text{synth}}, g)$. Following Eq. (7), the final predicted velocity is composed as:

$$\tilde{v}(x_t, t) = v_\theta(x_t, t, d_{\text{real}}, \emptyset) + \lambda\Big(v_\theta(x_t, t, d_{\text{synth}}, g) - v_\theta(x_t, t, d_{\text{synth}}, \emptyset)\Big) \tag{8}$$

This formulation elegantly uses the real branch to anchor the photorealistic appearance, while extracting a pure geometry correction term from the synthetic predictions. As a result, the framework successfully yields geometry-consistent Sim2Real synthesis without any paired supervision.

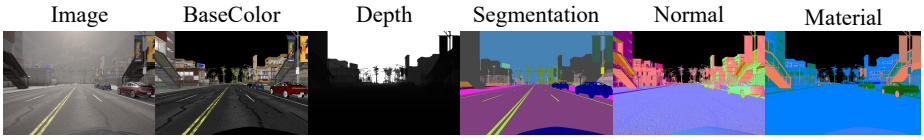


Fig. 3: Visualization of multi-modal geometry conditions. An original synthetic rendered image (left) and its corresponding G-buffers extracted via a deferred rendering pipeline.

3.3 Lightweight Multi-Condition Adapter

While the CDCT framework provides a principled mechanism for geometry-guided Sim2Real translation, directly adopting a standard ControlNet architecture leads to severe computational overhead. Conventional ControlNet replicates the entire encoder branch of the base model. For large Diffusion Transformers such as MM-DiT, especially under multiple dense geometric conditions, this duplication introduces prohibitive memory and computation costs.

A simple alternative is to append geometry condition tokens (e.g., G-buffers) to the input sequence. However, this approach also scales poorly. As the number of conditions increases, the attention cost grows quadratically, and features from different modalities become entangled.

To make the framework practical and scalable, the condition injection mechanism must be redesigned. In this subsection, we introduce the Lightweight Multi-Condition Adapter (LMCA), a parameter-efficient approach that injects multi-modal geometric guidance directly into attention layers. It avoids backbone duplication and reduces the complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$ as the number of conditions increases.

Geometry Conditions. We represent structured scene information using G-buffers derived from deferred rendering, including camera-space surface normals, depth, semantic segmentation, base color, and material buffers as shown in Fig. 3. Specifically, BaseColor $\in \mathbb{R}^{H\times W\times 3}$ corresponds to intrinsic albedo decoupled from illumination; Depth $\in \mathbb{R}^{H\times W\times 1}$ captures scene layout and inter-object spatial relationships; Segmentation $\in \mathbb{R}^{H\times W\times 1}$ provides object-level structural priors; Normal $\in \mathbb{R}^{H\times W\times 3}$ encodes local surface orientation and fine-scale geometric structure; Material $\in \mathbb{R}^{H\times W\times 3}$ contains metallic, specular, and roughness components in the RGB channels.

Condition Injection. To incorporate multiple geometry conditions without duplicating the MM-DiT backbone, we directly inject conditional signals through Low-Rank Adaptation (LoRA) [15]. Our approach reuses the frozen, pretrained attention projections and augments them with condition-specific LoRA weights as shown in Fig. 4. Rather than adding spatial residual features at the block outputs, we inject the geometric guidance strictly into the attention computation by projecting the condition embeddings into auxiliary key-value pairs.

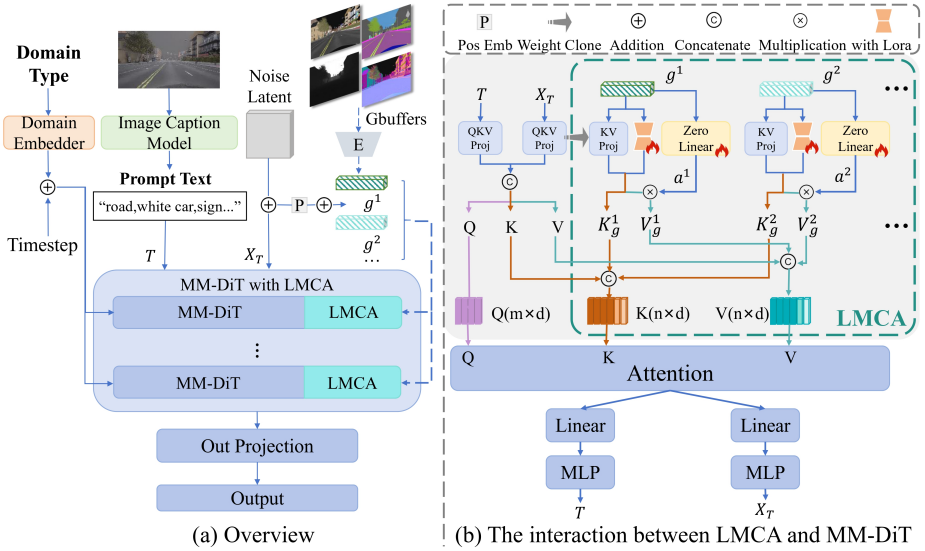


Fig. 4: Architecture of the MM-DiT with Lightweight Multi-Condition Adapter. Multiple conditions are processed through LoRA-augmented projections and a Zero Linear layer. The output scalar α scales the auxiliary **Value** component (V_g), while the **Key** component (K_g) remains unscaled. These conditional features are concatenated with the original keys and values (K, V) for the final attention computation. We extract the text prompt from the synthetic image using a large captioning model [2].

Concretely, for the i-th geometry condition, we generate condition-specific key and value features by applying low-rank adaptations to the pretrained attention projections. Formally, the conditional projections are defined as

$$K_g^i, V_g^i = (W_K + \Delta W_K^i)g^i, \quad (W_V + \Delta W_V^i)g^i. \tag{9}$$

where W_K, W_V are the shared projection matrices across image denoising branches. ΔW_K^i and ΔW_V^i are low-rank updates associated with the i-th condition, and g^i represents the embeddings corresponding to the i-th geometry condition. This formulation preserves the pretrained backbone structure while enabling condition-specific representation learning.

To ensure stable optimization while enabling adaptive multi-condition fusion, we introduce a zero-initialized linear layer for each condition:

$$\alpha^i = \text{ZeroLinear}(g^i), \quad V_g^i \leftarrow \alpha^i \cdot V_g^i \tag{10}$$

Crucially, only the Value component is modulated by the α^i, while the Key component K_g^i retains its projected structural information. This design ensures that at the start of training (where $\alpha^i = 0$), the condition has almost no contribution to the attention output, effectively preserving the pre-trained behavior of the base model. As training progresses, the layer learns non-zero scaling coefficients that implicitly allocate relative importance across different geometric modalities.

With the condition-specific key–value pairs constructed as above, we extend the original attention by concatenating all conditional keys and values with the pretrained text–noise tokens, and compute attention over the aggregated token set:

$$\text{softmax}\left(\frac{1}{\sqrt{d}}Q_{\text{x\&t}}[K_{\text{x\&t}}, K_g^1, \ldots, K_g^N]^\top\right)[V_{\text{x\&t}}, V_g^1, \ldots, V_g^N] \tag{11}$$

Under this formulation, each condition is incorporated as additional tokens within the same attention space rather than as a separate branch. The query $Q_{\text{x\&t}}$ attends jointly to the original text–noise keys and the condition-specific keys, with geometry guidance carried by the corresponding conditional values. This design enables efficient multi-conditional fusion while preserving structural disentanglement in large MM-DiT backbones, without auxiliary fusion modules or explicit cross-condition weighting.

4 Experiments

4.1 Experimental Setup

Datasets. To evaluate the effectiveness of our method and baseline methods in practical Sim2Real transfer, we construct a training setup consisting of synthetic source data and real target data. For synthetic data, we generate driving scenes using the CARLA simulator [7]. CARLA is a widely adopted open-source autonomous driving simulator that provides high-fidelity urban environments and precise control over scene configurations. Through its Python API, we render RGB images together with structured annotations, including semantic segmentation maps and G-buffers, which serve as geometric conditions in our framework. For real-world data, we use the Cityscapes dataset [6] due to its high-quality urban street imagery and strong compatibility with CARLA scene layouts. All semantic maps and G-buffers required by competing methods are obtained directly from the CARLA API to ensure consistency across baselines.

Implementation Details. We build our method upon Stable Diffusion 3 [8], and employ LENS [2] for image captioning. We utilize LoRA [15] with a default rank of 4 for the condition injection module. The training is performed in two stages: Stage-1 uses a learning rate of 1e-5, while Stage-2 adopts 1e-4. During inference, the sampling is applied with 25 steps and the guidance strength λ is set to 6—a value that strikes a good trade-off between fidelity, controllability, and image quality. All experiments are conducted on 4 NVIDIA L40S GPUs (48GB).

Baselines. For the comparison to the previous work, we select several generic unpaired image-to-image translation models that focus on better semantic alignment. Specifically, we compare our approach with several representative Sim2Real and unpaired image-to-image translation methods, including EPE [32], VSAIT [40],

and HIE [9], which focus on structural consistency in driving scenes, as well as unpaired image-to-image translation framework based on the Neural Schrödinger Bridge UNSB [19] and the diffusion-based CACTI_F [5].

Metrics. To evaluate Sim2Real performance, it is crucial to adopt appropriate and complementary criteria. We assess results from three dimensions: Controllability, Generative Quality, and Similarity. Controllability measures consistency between generated images and original input CARLA images. We use SSIM [42] to quantify structural alignment and LPIPS [48] to evaluate perceptual similarity in feature space. Generative Quality evaluates intrinsic image quality, focusing on artifacts, distortions, and perceptual naturalness. We employ NIQE [26] and MANIQA [43] as no-reference quality metrics. Similarity measures distribution alignment between translated images and Cityscapes dataset, reflecting the reduction of the Sim2Real gap across domains. We report FID [11] and KID($\times 10^3$) [3] to quantify cross-domain similarity at the feature-distribution level.

Additionally, we conduct downstream perception evaluation. For each method, we train Mask2Former models [4] for semantic segmentation, YOLOv8 models [17] for object detection, and NeWCRFs models [45] for monocular depth estimation on translated CARLA images with their corresponding annotations, and evaluate them on the Cityscapes validation set. We report mIoU for semantic segmentation and mAP_{50} for object detection. For depth estimation, we report AbsRel, SILog, $\text{RMSE}_{\log}$, and threshold accuracies under $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$. These metrics jointly assess whether translated synthetic data improves real-world perception across dense semantic understanding, object-level recognition, and geometry-sensitive prediction.

4.2 Comparison with Other Sim2Real Methods

As shown in Table 1, we compare our method against existing approaches across three dimensions. Our method consistently outperforms others in Generative Quality and Similarity, while maintaining competitive performance in Controllability. This indicates that our approach effectively reduces the Sim2Real gap

Table 1: Quantitative comparison with baseline methods on CARLA to Cityscapes.

Method	Controllability		Generative Quality		Similarity	
	SSIM↑	LPIPS↓	NIQE↓	MANIQA↑	FID↓	KID↓
EPE [32]	0.82	0.26	4.19	0.32	34.78	<u>25.23</u>
VSAIT [40]	**0.89**	<u>0.18</u>	<u>3.26</u>	<u>0.41</u>	<u>29.45</u>	27.67
HIE [9]	0.84	0.26	4.87	0.33	32.76	42.05
UNSB [19]	0.73	0.34	5.44	0.37	41.62	39.46
CACTI_F [5]	0.69	0.39	6.63	0.31	50.91	36.06
Ours	<u>0.85</u>	**0.17**	**3.21**	**0.44**	**26.45**	**22.17**

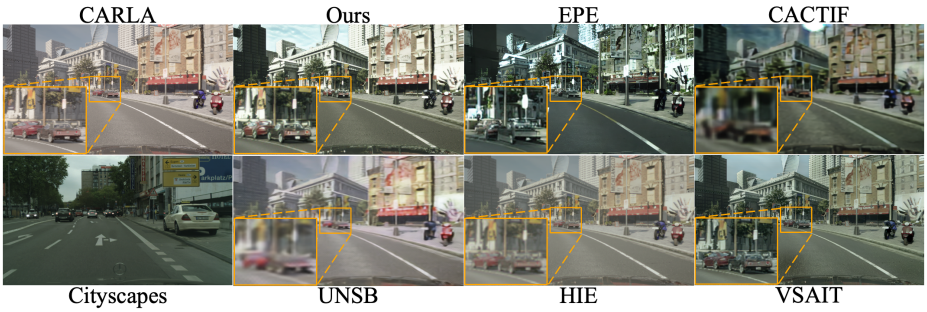


Fig. 5: Visual results of Sim2Real translation. Our approach synthesizes highly photorealistic appearances and better preserves structural details than baselines.

without sacrificing structural consistency with the source domain. Fig. 5 presents qualitative comparisons. Most methods are able to preserve the overall scene layout and major semantic structures. However, clear differences emerge in image quality. UNSB and CACTI_{F} tend to produce blurry results with limited high-resolution details. EPE often introduces noticeable artifacts and unstable textures. HIE shows inconsistencies in fine-grained structures and local appearance. VSAIT, although relatively stable, struggles to reproduce realistic glossiness and material properties. This is likely because its model does not incorporate explicit geometry or material-related conditional inputs, limiting its ability to capture fine-grained appearance variations. In contrast, our method generates sharper details, more natural textures, and improved material realism, while maintaining structural alignment, leading to more convincing Sim2Real translation results.

4.3 Sim2Real Methods for Perception Tasks

Semantic Segmentation. Table 2 shows that models trained on original CARLA data generalize poorly to Cityscapes, yielding near-random predictions for most categories. Although certain classes such as vegetation retain limited consistency, the overall performance indicates a severe Sim2Real gap. Compared with baseline methods, our translated data leads to the strongest overall segmentation performance, with consistent gains across major categories such as road, car, vegetation, and building. Some methods like UNSB and HIE improve specific classes yet remain limited by blurred details or unstable appearance, which affects dense prediction quality. However, a noticeable gap remains compared with models trained directly on Cityscapes. This discrepancy can be attributed to the inherent content gap [18] between CARLA and Cityscapes. The two datasets depict different geographic regions and traffic environments, leading to mismatches in region-specific categories such as traffic signs and traffic lights. In addition, certain object categories in CARLA exhibit limited intra-class diversity (e.g., pedestrians and buses), restricting generalization. By contrast, some classes are both region-invariant and well represented in CARLA. These include road, vege-

Table 2: Semantic segmentation results of Mask2Former models trained on original CARLA, translated CARLA, Cityscapes data and evaluated on the Cityscapes validation set.

Training Set	mIoU↑	Others	Road	Sidewalk	Person	Vegetation	Building	Car	Pole	Sky	Bus	T.sign	T.light
CARLA	14.6	11.9	22.1	1.8	1.6	40.2	27.9	36.1	3.2	40.7	4.7	2.9	2.0
EPE [32]	36.9	35.0	81.4	45.8	9.4	**72.3**	48.4	40.1	22.6	60.6	**26.4**	5.1	8.9
VSAIT [40]	33.8	29.8	78.9	40.0	18.3	59.6	43.5	48.2	22.1	**68.8**	8.1	7.9	8.3
HIE [9]	30.5	28.3	60.6	30.6	10.9	60.7	39.4	50.7	18.1	54.6	19.2	3.1	4.9
UNSB [19]	26.4	21.4	44.8	35.4	22.7	44.6	33.7	**55.4**	13.2	51.1	13.9	8.5	6.9
$CACTI_F$ [5]	32.1	27.3	74.2	60.1	16.5	51.1	46.4	38.4	20.6	54.9	12.6	7.0	10.1
Ours	**48.3**	**42.1**	**83.9**	**70.3**	**38.4**	68.7	**69.5**	**79.5**	**28.2**	60.3	24.9	**29.3**	**27.8**
Cityscapes	65.2	59.6	89.6	76.2	66.8	72.6	75.5	87.3	45.4	76.4	66.8	50.2	54.6

Table 3: Object Detection results of YOLOv8 models trained on original CARLA, translated CARLA, Cityscapes data and evaluated on the Cityscapes validation set.

Training Set	mAP_{50}↑	Person	Car	Truck	Bus	Motor.	Bicycle	Rider
CARLA	20.2	10.9	50.1	23.4	31.7	10.1	9.2	5.9
EPE [32]	33.9	21.3	65.2	43.4	40.3	30.4	24.6	11.8
VSAIT [40]	35.7	29.7	67.8	39.2	37.8	**39.7**	18.9	18.5
HIE [9]	31.8	24.9	63.2	37.4	36.2	33.7	16.4	10.8
UNSB [19]	28.6	13.5	58.4	30.7	38.1	34.5	15.9	9.1
$CACTI_F$ [5]	31.7	22.5	57.4	39.7	36.4	29.9	19.0	16.7
Ours	**40.5**	**36.9**	**70.1**	**45.2**	**44.8**	38.5	**29.4**	**18.6**
Cityscapes	48.8	48.1	85.3	50.6	52.3	43.1	38.8	23.7

tation, and car. Such classes benefit most from the translation and can approach real-data performance.

Object Detection. A similar trend is observed in Table 3. Training on raw CARLA data results in limited detection performance on real images. Translated data consistently improves detection accuracy across most categories, especially for common traffic participants (e.g., cars). This confirms that narrowing the appearance gap leads to better cross-domain detection performance. Nevertheless, the remaining gap to Cityscapes-trained models again reflects the content gap.

Depth Estimation. Table 4 reports the monocular depth estimation results of NeWCRFs models trained on different translated datasets and evaluated on the Cityscapes validation set. Training directly on raw CARLA images leads to poor

Table 4: Depth estimation results of NeWCRFs models on the Cityscapes validation set.

Training Set	AbsRel↓	SILog↓	$RMSE_{log}$↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
CARLA	0.56	44.18	0.53	0.29	0.55	0.77
EPE [32]	0.20	25.04	<u>0.24</u>	0.73	0.84	0.91
VSAIT [40]	<u>0.19</u>	21.15	<u>0.24</u>	<u>0.78</u>	<u>0.85</u>	<u>0.92</u>
HIE [9]	0.21	28.93	0.25	0.68	0.77	0.85
UNSB [19]	0.25	39.11	0.27	0.51	0.67	0.79
$CACTI_F$ [5]	0.23	**18.06**	0.26	0.60	0.71	0.81
Ours	**0.16**	<u>19.24</u>	**0.23**	**0.81**	**0.91**	**0.96**
Cityscapes	0.11	14.92	0.19	0.87	0.96	0.99

Table 5: Impact of training data composition on object detection and semantic segmentation performance.

Training Set	mAP_{50}↑	mIoU↑
Cityscapes	48.8	65.2
CARLA	20.2	14.6
Ours	40.5	48.3
CARLA + Cityscapes	38.9	49.8
Ours + Cityscapes	**52.5**	**67.0**

cross-domain generalization, with high AbsRel and SILog errors, indicating that the synthetic-to-real appearance gap also substantially affects geometry-sensitive perception tasks.

Our method achieves the best overall performance among translated-data methods, obtaining the lowest AbsRel of 0.16, the lowest $RMSE_{log}$ of 0.23, and the highest accuracy under all three threshold metrics. Although $CACTI_F$ obtains slightly lower SILog, its threshold accuracies remain notably lower than ours, indicating less stable depth prediction across pixels.

Combined Data Training. Table 5 highlights the practical value of Sim2Real translation. Although translated data alone does not match performance achieved by training solely on real data, combining translated CARLA data with Cityscapes yields further improvements over using real data alone. This indicates that translated synthetic data provides complementary diversity and scale, serving as an effective data augmentation strategy for real-world autonomous driving perception tasks.

4.4 Ablation Study

Effect of Cross-Domain Score Composition. Table 6 studies the effect of the proposed score composition mechanism. In this ablation, we remove score composition and directly feed the geometry conditions together with the real-domain indicator d_{real} during inference. As shown in Fig. 6, without score composition, the model tends to prioritize satisfying the G-buffers constraints while

Table 6: Ablation study on the effect of Score Composition in the data generation pipeline.

Ablation	FID↓	KID↓
w/o Score Comp.	63.77	57.11
w/ Score Comp.	**26.45**	**22.17**

Table 7: Parameter overhead compared with standard ControlNet (M.Cond.: multi-condition support; N: number of conditions).

Method	M.Cond.	Params
ControlNet [47]	✘	1.1B (+55 %)
Ours	✔	57M (+2.8 %)×N

Fig. 6: Qualitative examples of Score Composition.

largely ignoring the real-domain appearance prior. In contrast, our cross-domain score composition explicitly combines the real-domain appearance prior with the geometry correction term extracted from the synthetic domain, enabling geometry-consistent synthesis while effectively improving realism.

Effect of Lightweight Multi-Condition Adapter. Table 7 compares the parameter overhead of our LMCA with standard ControlNet. Conventional ControlNet requires duplicating a large portion of the backbone, resulting in substantial parameter growth. In contrast, our design injects multiple conditions through lightweight LoRA-based projections within the attention layers. This introduces only a small number of additional parameters and minimal computational overhead, both scaling linearly with the number of conditions. This validates that direct key–value augmentation within the attention space provides a more scalable and computation-efficient alternative while maintaining performance.

5 Conclusion

In this paper, we presented Cross-Domain Control Transfer (CDCT), a novel diffusion-based Sim2Real translation framework tailored for autonomous driving. This framework combines cross-domain score composition and LMCA to ensure unpaired spatial consistency with minimal parameter overhead. Extensive CARLA-to-Cityscapes evaluations validate its superiority. CDCT not only delivers substantial zero-shot performance improvements for downstream perception models but also demonstrates that joint training with our generated data can successfully surpass the performance of models trained exclusively on real-world datasets.

While our method effectively narrows the appearance gap, the inherent content gap, such as discrepancies in scene layouts and object distributions, remains an open challenge.

References

1. Abu Alhaija, H., Mustikovela, S.K., Geiger, A., Rother, C.: Geometric image synthesis. In: Asian Conference on Computer Vision. pp. 85–100. Springer (2018)
2. Berrios, W., Mittal, G., Thrush, T., Kiela, D., Singh, A.: Towards language models that can see: Computer vision through the lens of natural language. arXiv preprint arXiv:2306.16410 (2023)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
4. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
5. Chigot, E., Wilson, D.G., Ghrib, M., Oberlin, T.: Style transfer with diffusion models for synthetic-to-real domain adaptation. Computer Vision and Image Understanding p. 104445 (2025)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
7. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
9. Fan, W., Chen, J., Liu, Z.: Hierarchy flow for high-fidelity image-to-image translation. arXiv preprint arXiv:2308.06909 (2023)
10. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
14. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. pp. 1989–1998. Pmlr (2018)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
16. Jaunet, T., Bono, G., Vuillemot, R., Wolf, C.: Sim2realviz: Visualizing the sim2real gap in robot ego-pose estimation. arXiv preprint arXiv:2109.11801 (2021)
17. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 (2023), `https://github.com/ultralytics/ultralytics`
18. Kar, A., Prakash, A., Liu, M.Y., Cameracci, E., Yuan, J., Rusiniak, M., Acuna, D., Torralba, A., Fidler, S.: Meta-sim: Learning to generate synthetic datasets. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4551–4560 (2019)

19. Kim, B., Kwon, G., Kim, K., Ye, J.C.: Unpaired image-to-image translation via neural schrödinger bridge. In: International Conference on Learning Representations. vol. 2024, pp. 19312–19331 (2024)
20. Kobyzev, I., Prince, S.J., Brubaker, M.A.: Normalizing flows: An introduction and review of current methods. IEEE transactions on pattern analysis and machine intelligence **43**(11), 3964–3979 (2020)
21. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
22. Li, M., Yang, T., Kuang, H., Wu, J., Wang, Z., Xiao, X., Chen, C.: Controlnet++: Improving conditional controls with efficient consistency feedback: Project page: liming-ai. github. io/controlnet_plus_plus. In: European Conference on Computer Vision. pp. 129–147. Springer (2024)
23. Li, Y., Wu, Z., Zhao, H., Yang, T., Liu, Z., Shu, P., Sun, J., Parasuraman, R., Liu, T.: Aldm-grasping: diffusion-aided zero-shot sim-to-real transfer for robot grasping. arXiv preprint arXiv:2403.11459 (2024)
24. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
25. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
26. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
27. Mo, S., Mu, F., Lin, K.H., Liu, Y., Guan, B., Li, Y., Zhou, B.: Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7465–7475 (2024)
28. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
29. Pahk, J., Shim, J., Baek, M., Lim, Y., Choi, G.: Effects of sim2real image translation via dclgan on lane keeping assist system in carla simulator. IEEE Access **11**, 33915–33927 (2023)
30. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. pp. 319–345. Springer (2020)
31. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
32. Richter, S.R., AlHaija, H.A., Koltun, V.: Enhancing photorealism enhancement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 1700–1715 (2022)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
35. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
36. Song, Y., Lou, S., Liu, X., Ci, H., Yang, P., Liu, J., Shou, M.Z.: Anti-reference: Universal and immediate defense against reference-based generation. arXiv preprint arXiv:2412.05980 (2024)

37. Song, Y., Yang, P., Ci, H., Shou, M.Z.: Idprotector: An adversarial noise encoder to protect against id-preserving image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3019–3028 (2025)
38. Su, X., Song, J., Meng, C., Ermon, S.: Dual diffusion implicit bridges for image-to-image translation. arXiv preprint arXiv:2203.08382 (2022)
39. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14940–14950 (2025)
40. Theiss, J., Leverett, J., Kim, D., Prakash, A.: Unpaired image translation via vector symbolic architectures. In: European Conference on Computer Vision. pp. 17–32. Springer (2022)
41. Wang, H., Peng, J., He, Q., Yang, H., Jin, Y., Wu, J., Hu, X., Pan, Y., Gan, Z., Chi, M., et al.: Unicombine: Unified multi-conditional combination with diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18325–18334 (2025)
42. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
43. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1191–1200 (2022)
44. Yu, A., Foote, A., Mooney, R., Martín-Martín, R.: Natural language can help bridge the sim2real gap. arXiv preprint arXiv:2405.10020 (2024)
45. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: Neural window fully-connected crfs for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3916–3925 (2022)
46. Zhang, C.Y., Shrivastava, A.: Aptsim2real: Approximately-paired sim-to-real image translation. arXiv preprint arXiv:2303.12704 (2023)
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
49. Zhang, Y., Yuan, Y., Song, Y., Wang, H., Liu, J.: Easycontrol: Adding efficient and flexible control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19513–19524 (2025)
50. Zhao, H., Wang, Y., Bashford-Rogers, T., Donzella, V., Debattista, K.: Exploring generative ai for sim2real in driving data synthesis. In: 2024 IEEE Intelligent Vehicles Symposium (IV). pp. 3071–3077. IEEE (2024)
51. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Unicontrolnet: All-in-one control to text-to-image diffusion models. Advances in neural information processing systems **36**, 11127–11150 (2023)
52. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
53. Zou, S., Huang, Y., Yi, R., Zhu, C., Xu, K.: Cyclediff: Cycle diffusion models for unpaired image-to-image translation. IEEE Transactions on Image Processing (2026)