AM-Adapter: Appearance Matching Adapter for Exemplar-based Semantic Image Synthesis in-the-Wild

Supplementary Material

In this material, Section A describes the implementation details of our experiments, while Section B elaborates the details of the evaluation process. Section C presents additional results, including qualitative, quantitative analyses and user studies. Finally, in Section D, we demonstrate the versatility of our work through various applications. Section E discusses the limitations of our work and provides a broader discussion.

A. Implementation Details

For all experiment, we used a single NVIDIA A6000 GPU. For training, we utilized 7K exemplar-target segmentation pairs from the BDD100K [32] dataset. During the first stage training, ControlNeXt [23] was trained with a batch size of 2 and a resolution of 512, using a learning rate of $1e-5$, and the training process lasted approximately 9 hours. In the second stage, AM-Adapter was trained with a batch size of 1 and a resolution of 512 for $25,000$ steps, taking around 14 hours. We used the same learning rate $1e-5$, with the AdamW [15] optimizer applied in both stages. We utilized the DDIM [26] sampler for both inversion and sampling, with the number of timesteps T set to 20. The Augmented Self-Attention is applied across all self-attention layers of the UNet, denoted as $L \in [0, 9]$, while the AM-Adapter is applied to all self-attention layers except for the first block of the encoder, represented as $L \in [1, 9]$. This approach avoids the structural conflict and prevents interference between the two effects, as the cross-normalized features from ControlNeXt [23] are added after the first downsampling block. We set the guidance scale to $s = 7.5$ for both text and matching guidance.

B. Evaluation

B.1. Dataset

Prior works in Semantic Image Synthesis [10, 11, 18, 21, 23, 29, 33, 34] and Exemplar-based Semantic Image Synthesis [6, 12, 24, 30, 31] have used different datasets for evaluation. Commonly, prior methods [13] assemble image dataset from the web and hand annotated the condition such as segmentation, sketch or edge drawing. Since no established benchmark exists for Exemplar-based Semantic Image Synthesis, we constructed an evaluation dataset tailored to our task, featuring complex driving scenes with diverse structure-appearance pairs. Specifically, we evaluated our method on two commonly used driving scene datasets: BDD100K [32] and Cityscapes [4]. To demonstrate generalization ability, we additionally evaluated our method on the NYUv2 [20] dataset, which is an indoor dataset. For evaluation, we randomly selected 300 segmentation maps each from BDD100K Cityscapes and NYUv2, resulting in a total of 900 segmentation maps.

Our goal is to achieve semantic-aware local appearance transfer in complex scenarios, which selectively transfers exemplar's local appearance to target segmentation under significant geometric gap between the exemplar's and target's layouts, and with multiple instances. As part of this effort, we propose a retrieval technique that automatically selects exemplars while maximizing matchable regions, as discussed in Section 3.6 of the main paper. Therefore, the exemplar segmentation-image pairs in our evaluation dataset consist of two types: 300 pairs retrieved using 300 target segmentation maps and 300 pairs randomly selected.

B.2. Evaluation Metrics

For quantitative evaluation, we categorized our analysis into three perspectives: structure consistency, appearance preservation and image quality. Following previous studies [13, 22], we adopted Self-Sim. [28] metrics to evaluate structure consistency. To evaluate appearance preservation, following prior works [19], we computed CLIP image similarity [25], denoted as I_{CLIP}. For image quality assessment, we measured FID [5]. Table 1 of the main paper presents the results for these metrics.

While our AM-Adapter outperforms other methods across all metrics, as noted in [3, 8, 9, 14, 27], we emphasize that these metrics do not fully align with human preferences. This is because they extract either global features or small local features from the generated images and conditions (segmentation maps and exemplar RGB images) and calculate distances between these features. To address this limitation, we conducted the user study for a more reliable evaluation.

Furthermore, we evaluated three complementary metrics to address this limitation: object-wise local CLIP similarity, I_{DINO} and DINO [cls] loss. The object-wise local CLIP similarity is computed by categorizing objects and measuring the CLIP image similarity for each category individually. Since meaningful comparison requires the presence of corresponding objects in both the exemplar and result images, we restrict our analysis to the top 10 most frequently occurring object classes in the BDD100K [32] dataset. Despite this constraint, our method consistently demonstrates superior robustness across all classes compared to other

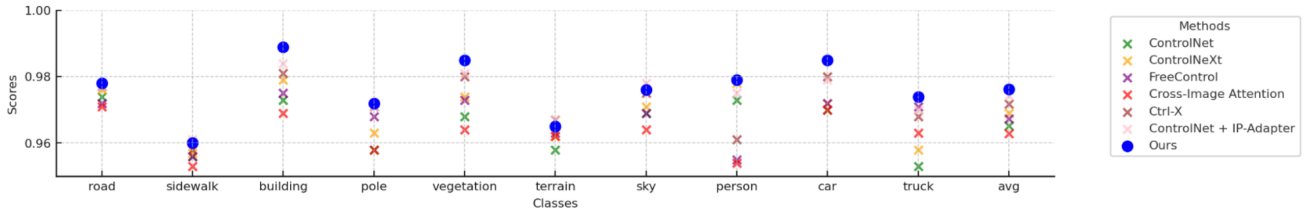


Figure 1. **Object-wise CLIP Image Similarity per Class.**

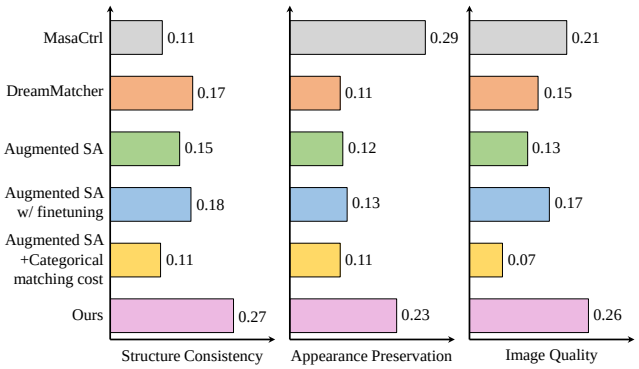


Figure 2. **User Study on Ablations.**

baseline models. Additionally, we employ DINO, which is known for its supervised learning capability and its ability to capture fine-grained local details, to measure image similarity. Our approach achieves the best performance in this evaluation. Table 1 of the main paper demonstrates the results for I_{DINO} and DINO [cls] loss, and Figure 1 shows the results of object-wise local CLIP similarity across categories in BDD100K dataset.

B.3. User Study Details

Figure 6 presents example questions from the user study. We conducted a human evaluation study comparing AM-Adapter and previous works [1, 13, 17, 31, 33] in terms of structure consistency, appearance preservation, and image quality. For structure consistency, we provided the target segmentation map and the generated images from different methods, and users were asked to select which method better represents the semantic structure in the target segmentation map. For appearance preservation, we provided the exemplar image and the generated images from different methods, and participants were asked to choose which generated image better captures the appearance in the exemplar. Lastly, participants were shown only the generated images and asked to select the one that achieved the highest image quality. A total of 45 participants responded to 18 questions. For a fair comparison, we sampled generated images from a large pool sharing the same exemplar image for three different methods to ensure intra-rater reliability. Figure 9 in the main paper summarizes the results, showing that our model outperforms others across all three criteria.

Figure 7 shows example questions from the user study comparing AM-Adapter with its individual components. The evaluation was conducted using the same criteria of structure preservation, appearance preservation and image quality to assess the role of each component in the overall performance. A total of 33 participants responded to 24 questions. For a fair comparison, generated images were samples from a shared pool associated with the same exemplar image and same target segmentation map across all methods. The results, summarized in Figure 2, show that AM-Adapter achieves a balanced and consistently high performance across all three criteria.

C. Additional Results

C.1. Qualitative Results

Figure 8 shows additional qualitative results of AM-Adapter. We present more qualitative results comparing our method with others, including ControlNet [33] + IP-Adapter [31], FreeControl [17], Cross-Image Attention [1], and Ctrl-X [13], in Figure 9, which further demonstrates the effectiveness of our method.

C.2. Ablation Study Analysis

Table 2 in the main paper, Figure 3, Figure 4 and Table 1 in Appendix summarize the following ablation study.

MasaCtrl vs. DreamMatcher vs. Augmented Self-Attention. As shown in Figure 3 and Table 2, we conducted a comparative analysis of the Augmented Self-Attention against two representative hand-crafted attention control methods, MasaCtrl [2] and DreamMatcher [19], which focus on implicit matching within self-attention mechanisms. MasaCtrl replaces the key and value of the synthesized target image with those of the exemplar image, while DreamMatcher enhances implicit matching by leveraging diffusion features for improved correspondence.

MasaCtrl heavily relies on implicit matching in the self-attention module, often resulting in inaccurate appearance transfers to semantically misaligned regions, thereby disrupting the target structure that aligns with the structural conditions. Furthermore, key-value replacement cannot generate new elements that are absent in the exemplar, as it discards the original keys and values from the target and replaces them with those from the exemplar. This limitation causes key-value replacement (MasaCtrl) to exhibit

Figure 3. **Ablation Study on Individual Components:** (a) ControlNeXt [23] + MasaCtrl [2], (b) ControlNeXt [23] + DreamMatcher [19], (c) ControlNeXt [23] + Augmented Self-Attention, (d) (c) + w/ Fine-tuning, (e) (c) + Categorical Matching Cost, (f) **AM-Adapter (Ours)**.

high I_{CLIP} and I_{DINO} and low DINO [cls] loss, as it discards the original key-value pairs of the target and copies and pastes the exemplar's appearance into the target image based on incorrect matching. This is further demonstrated by the Self-Sim. [28] of MasaCtrl in Table 2 and the structural consistency results from the user study Figure 2, which show that key-value replacement disrupts the structural consistency between generated images and the given semantic conditions. Notably, the metric of exemplar image in Table 2 reflects cases where the same appearance exemplar is used, which naturally leads to very high I_{CLIP} scores due to the identical appearance. Similarly, MasaCtrl achieves high scores even when matching fails, as it transfers all appearance information from the exemplar to the target.

In contrast, as demonstrated in Figure 3 and Table 2, DreamMatcher better preserves the target structure by leveraging improved semantic matching and using intermediate diffusion features to align the exemplar's appearance with the target image. However, its appearance preservation is highly dependent on the exemplar image, as relying solely on hand-crafted matching with diffusion features is insufficient for establishing accurate correspondence in content-rich images, such as driving scenes or indoor scenes, which require precise matching.

As a result, to achieve our goal of accurate transfer in complex scenes, we adopt Augmented Self-Attention as our baseline, which is outlined in Section 3.3 of the main paper. As shown in Figure 3 and Table 2, Augmented Self-Attention selectively incorporates keys and values from both the exemplar and the target, effectively preserving the desired target structure while transferring the exemplar's local appearance. Due to this selectivity, I_{CLIP} is lower compared to key-value replacement, which indiscriminately transfers the entire appearance information from the exemplar. However, Augmented Self-Attention demonstrates a superior human evaluation score compared to MasaCtrl [2] and DreamMatcher [19], indicating that it concurrently achieves appearance transfer and structural consistency.

Impact of Fine-tuning. Nevertheless, as discussed in Section 3.3 in the main paper, relying solely on the implicit matching of the Augmented Self-Attention still leads to the mismatches. The most straightforward approach to address this issue is fine-tuning the model. However, as shown in Figure 3 (c) and (d) and discussed in Section 4.3, it is shown that fine-tuning degrades detailed appearance preservation, as it requires learning both generation and matching simultaneously, which leads to unstable training and overfitting. This is further demonstrated in the (V) in Table 2, (d) in Figure 3.

Categorical Matching Cost vs. Learnable Matching Cost (AM-Adapter). To concurrently achieve matching and generation, we refine the implicit matching with semantic awareness using segmentation maps in a data-driven manner, rather than fine-tuning the entire model. In Section 3.4 of the main paper, we compare categorical matching cost with AM-Adapter. (e) and (f) in Figure 3 further demonstrate the effectiveness of AM-Adapter compared to categorical matching cost. The algorithm of overall AM-Adapter training is available in Algorithm 1.

Inference. Figure 4 and Table 1 illustrate the effects of our proposed retrieval technique and matching guidance during inference. (a) represents a randomly selected exemplar image, (b) represents a retrieved exemplar image, and (c) is the target segmentation map with desired structure. (d) and (I) display the results when a random exemplar image is used as input without retrieval. In contrast, (e) and (II) utilize an exemplar image that is structurally most similar to the target condition, resulting in higher appearance preservation compared to (d). Notably, this retrieval-based approach results in a significant increase in I_{CLIP} scores, increasing the matchable regions and transferring more appearance information from the exemplar. Finally, (f) and (III) show the results of applying matching guidance using classifier-free guidance [7], highlighting the effectiveness of matching guidance in achieving accurate results. Figure 11 presents the detailed mechanism of how the retrieval process operates and Figure 12 illustrates retrieved exemplar images from target segmentation maps obtained by our retrieval technique. The algorithm of overall AM-Adapter inference is available in Algorithm 2.

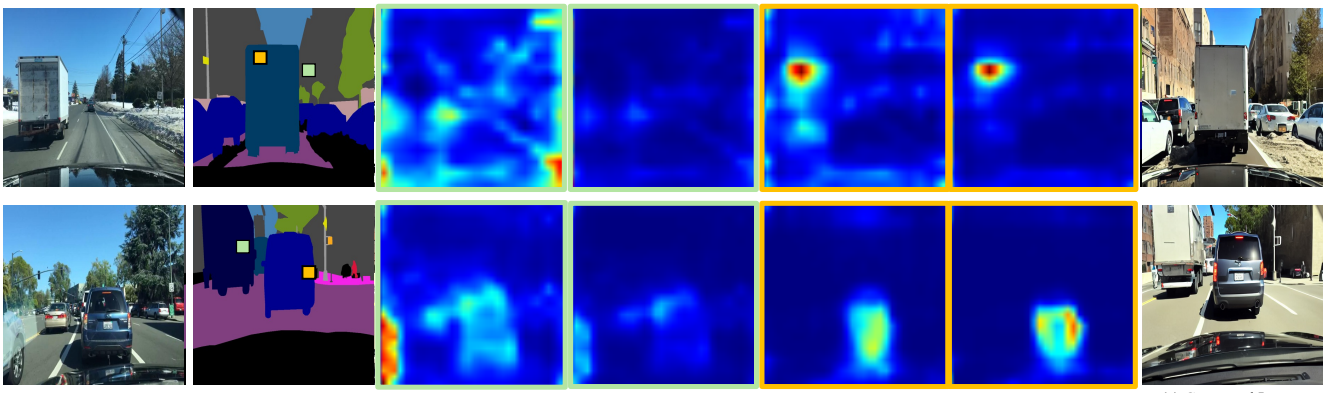
C.3. Additional Analysis

Figure 5 presents the additional examples of attention visualizations before and after applying the AM-Adapter, further highlighting the effectiveness of our model. The green markers indicate objects that are absent in the exemplar, while the orange markers denote objects that are present in the exemplar.

For instance, in the first row, the green marker highlights a query point corresponding to a 'building' that is absent in the exemplar. After applying the AM-Adapter, the attention associated with unrelated regions is significantly suppressed. The orange marker, on the other hand, indicates a query point on the top of a 'trailer'. After applying AM-Adapter, the attention becomes more localized, concentrating more effectively on the relevant regions. In the second row, the green marker represents a query point corresponding to a 'truck' that does not appear in the exemplar. After the AM-Adapter is applied, mismatches in the attention are reduced, further aligning the attention with the target structure. Similarly, the orange marker identifies a query point near the right rear light of a 'vehicle'. With the adapter applied, the attention becomes more localized, exhibiting enhanced focus on the intended area.

(a) (b) (c) (d) (e) (f)

Figure 4. **Ablation on Inference:** (a) random exemplar image, (b) retrieved exemplar image, (c) target segmentation map with desired structure, results generated (d) without retrieval or matching guidance, (e) with retrieval but without matching guidance, and (f) with both retrieval and matching guidance.



(a) Exemplar Image (b) Target Segmentation (c) Attention w/o Adapter (d) Attention w/ Adapter (e) Attention w/o Adapter (f) Attention w/ Adapter (e) Generated Image

Figure 5. **Additional Attention Visualization:** (a) Exemplar image with desired appearance, (b) target segmentation with desired structure, and (g) generated image. Green and orange markers in (b) indicate query points. (c) and (d) show the augmented self-attention map $Q_t^Y (K_t^X)^T$ from the green marker, before and after applying AM-Adapter, respectively. (e) and (f) show the augmented self-attention map $Q_t^Y (K_t^X)^T$ from the orange marker, before and after applying AM-Adapter, respectively.

	Component	Self-Sim. ↓	I_{CLIP} ↑	FID↓
(I)	AM-Adapter	0.043	0.741	79.05
(II)	(I) + Retrieval	<u>0.041</u>	<u>0.814</u>	<u>77.15</u>
(III)	(II) + Matching Guidance **(Ours)**	**0.041**	**0.819**	**75.89**

Table 1. **Ablation Study on Inference.**

C.4. Generalization

Our method is not limited to driving scenes. To demonstrate its generalizability across domains, we additionally evaluated it on NYUv2 [20], an indoor dataset with 40 classes, as shown in Figure 10.

D. Applications

D.1. Controllable One-to-One Appearance Transfer with User Guidance

As illustrated in Figure 13, the AM-Adapter can enforce one-to-one mapping within a many-to-many setting, allowing precise control over appearance transfer. In the user-defined exemplar segmentation, the white regions indicate source vehicles whose appearance is explicitly designated for transfer. In the user-defined target segmentation, the white regions represent destination objects that will receive the transferred appearance. As a result, the appearance of the selected source vehicle in the exemplar is accurately mapped to the corresponding destination vehicle in the target segmentation.

D.2. Segmentation-Guided Image Editing

Figure 14 and Figure 15 illustrate the results of AM-Adapter to segmentation-based editing applications, specificallly object removal and addition, respectively. These results showcase the versatility of AM-Adapter in generating high-quality images while adhering to modified semantic structures.

In Figure 14, we demonstrate segmentation-based editing by removing objects from the segmentation maps. (a) represents the original segmentation map, while (b) shows the generated image by AM-Adapter following the target structure of (a). (c) and (e) are edited segmentation maps derived from (a) with specific modifications, such as modifying or removing the colors of specific instances within the map to adjust the semantic information. For example, in the first row, (c) removes the vehicles in the center, while (e) removes the buildings on the left. In the second row, (c) eliminates the traffic lights and trees, while (e) removes the person in the enter. (d) is the generated image by AM-Adapter following the target structure of (c), while (f) is the generated image by AM-Adapter following the target structure of (e).

Figure 15, in contrast, showcases segmentation-based editing through object addition in the segmentation maps. (a) depicts the edited segmentation maps from Figure 14 where specific objects were removed, forming the basis for subsequent augmentation, and (b) presents the images generated based on (a). (c) and (e) illustrate segmentation maps augmented by introducing additional semantic instances to enrich the scene. For example, in the first row, (c) introduces pedestrians and trees into the left side of the scene, while (e) adds buildings to the background. In the second row, (c) incorporates an additional vehicle into the scene, while (e) includes a pedestrian next to the existing one. The resulting images, generated based on these augmented segmentation maps, are depicted in (d) and (f), highlighting the model's capability to seamlessly integrate new objects into the scene while preserving global consistency.

The results shown in Figure 14 and Figure 15 highlight the versatility of AM-Adapter in addressing diverse semantic editing tasks, including object removal and addition. The method effectively synthesizes images that reflect both the structural modifications and the appearance of exemplar, showcasing its potential for downstream applications such as semantic image editing [16]. These findings further validate AM-Adapter's contribution to advancing exemplar-based semantic image synthesis.

D.3. Image-to-Image Translation

In Figure 16, we present the results of AM-Adapter in image-to-image translation. The first and third rows showcase exemplar images representing a diverse range of weather conditions and times of day. The second and fourth rows illustrate the generated results, where the detailed appearance of the exemplar image is seamlessly transferred into the desired structure of the given target segmentation map. Notably, even within the same category, subtle variations exist. For instance, sunny conditions can vary between partly cloudy and completely clear skies, the sky can display different hues during sunset, and night scenes can have varying levels of brightness. Describing such nuanced differences using textual descriptions is inherently challenging. By using exemplar images to replace ambiguous textual descriptions, AM-Adapter can translate the appearance of the given image into the user-intended local exemplar appearance, with segmentation maps serving as anchors.

D.4. Appearance-Consistent Consecutive Video Frame Generation

As illustrated in Figure 17, AM-Adapter effectively generates appearance-consistent consecutive video frames when provided with the target segmentation maps for each frame and a single exemplar image. The first row depicts the target segmentation maps of consecutive frames provided by the BDD100K [32] dataset, arranged in temporal order from left to right, while the second row shows the corresponding generated images that reflect the structure of each target map.

It is critical to note that, as mentioned in Section 3.5, our method was trained using pairs of images generated by applying random augmentations, such as flipping and cropping, to a single anchor image. This training process was designed to facilitate local appearance transfer, without explicitly considering inter-frame continuity or consistency. Nevertheless, our method effectively transfers the appearance of the exemplar image, resulting in consecutive frames that maintain appearance consistency. These results indicate that the performance of our model could be further refined by fine-tuning with video-image and segmentation pairs explicitly designed to ensure temporal consistency.

E. Limitation and Discussion

Dependency on Pretraining. As discussed in Section 3.5, our training process is divided into two stages: ControlNeXt [23] for generation and AM-Adapter for matching. In the first stage, ControlNeXt learns to generate realistic images using segmentation maps as conditions. However, if ControlNeXt fails to adequately learn the generative capabilities or fully grasp the semantic information from the segmentation maps, it can negatively affect the performance of the learnable matching cost in AM-Adapter, since it is learned in a data-driven manner using segmentation-image pairs. This underscores the importance of robust pretraining for ControlNeXt.

Limited Temporal Consistency Under Large Scene Changes. In Figure 17 and Section D.4, we demonstrated

our method's ability to generate appearance-consistent frames using an exemplar image and the target segmentation maps of consecutive video frames as input. However, when there are substantial scene changes (e.g., large camera motion), the consistency between generated frames diminishes.

As mentioned in Section D.4, this limitation arises from our training approach, which uses pairs of randomly augmented images derived from a single anchor image. While this method effectively accounts for spatial differences within a pair of frames, it does not address temporal consistency across frames, resulting in temporally inconsistent images during large scene transitions. To overcome this limitation, fine-tuning the adapter with video data that explicitly incorporates temporal consistency during training could enhance its ability to generate consistent frames, even under significant scene variations.

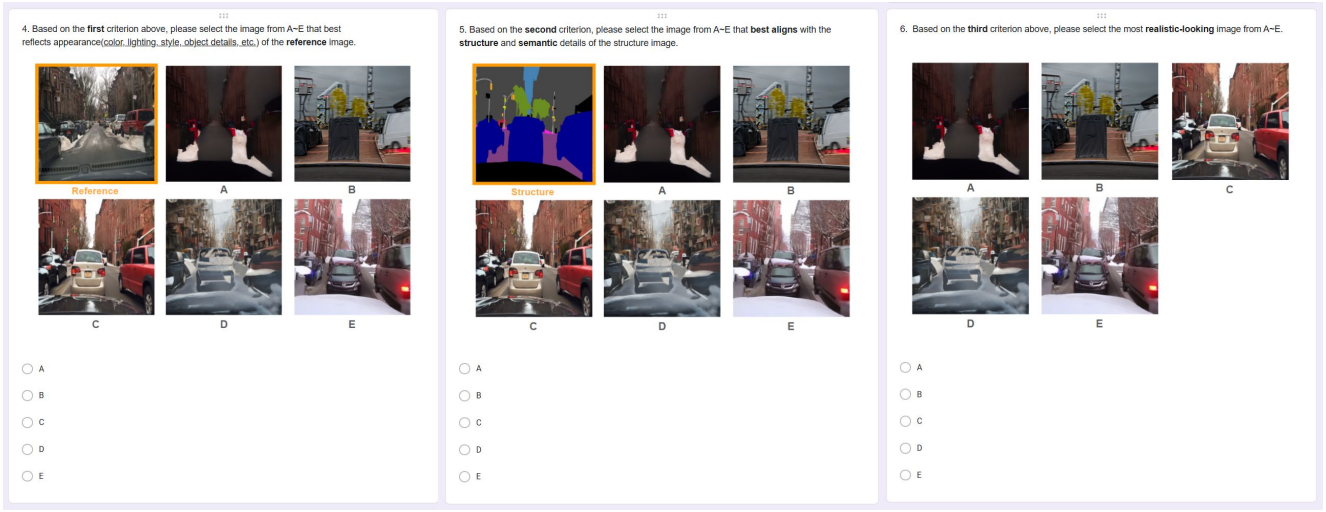


Figure 6. **An Example of a User Study Comparing AM-Adapter with Previous Methods.** For structure preservation, we provide the target segmentation map and generated images from different methods, ControlNet [33] + IP-Adapter [31], FreeControl [17], Cross-Image Attention [1], Ctrl-X [13] and AM-Adapter. For appearance preservation, we provide the exemplar image and the generated images from those methods. For image quality, we compare solely the generated images. For a fair comparison, we randomly select the sample generated from exemplar-segmentation pairs from a large pool.

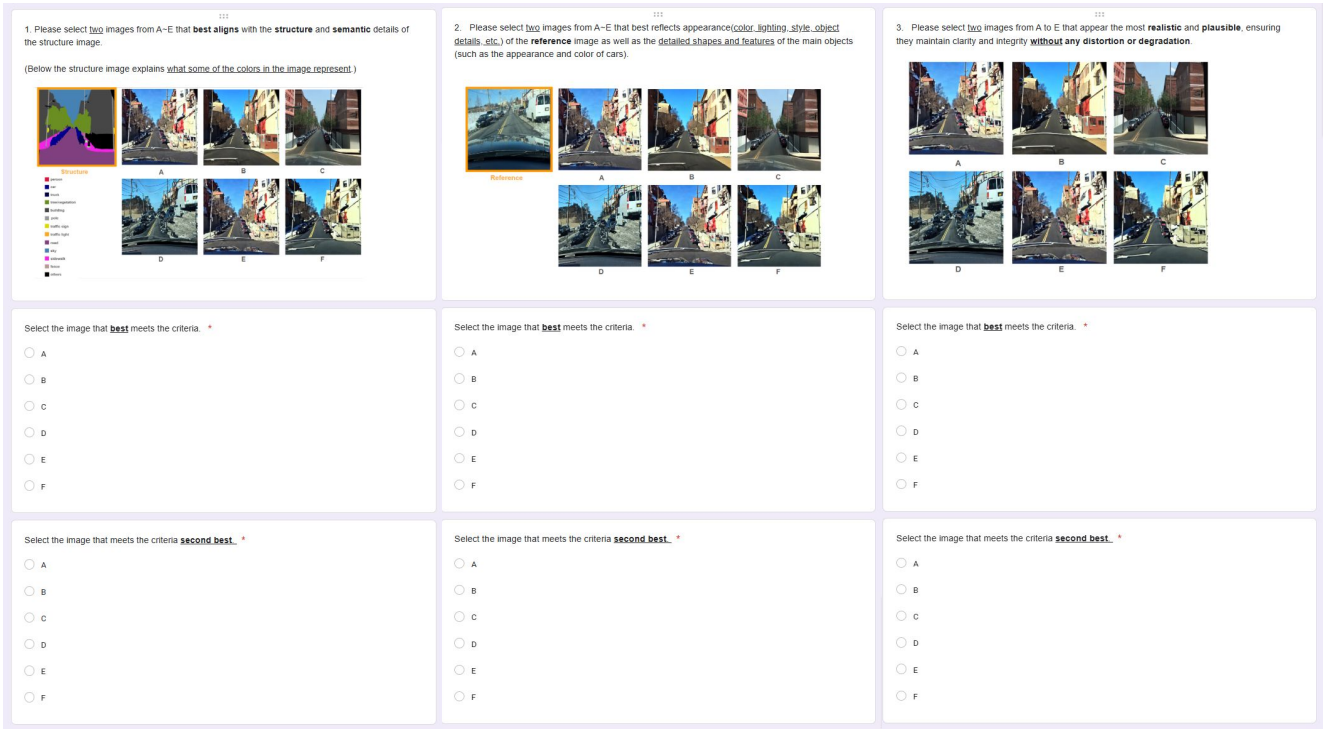


Figure 7. **An Example of a User Study Comparing AM-Adapter with Ablation Studies.** For structure preservation, we provide the target segmentation map and generated images from different methods, key-value replacement (MasaCtrl [2]), DreamMatcher [19], augmented self-attention, augmented self-attention w/ finetuning, augmented self-attention + categorical matching cost, and **AM-Adapter**. For appearance preservation, we provide the exemplar image and the generated images from those methods. For image quality, we compare solely the generated images. For a fair comparison, we randomly select the sample generated from exemplar-segmentation pairs from a large pool.

Algorithm 1 Training AM-Adapter with Frozen ControlNeXt and Diffusion

Input: Exemplar and target latent maps z_t^X, z_t^Y, segmentation maps S^X, S^Y, ground-truth noise ϵ
Output: Trained AM-Adapter parameters ϕ

Freeze parameters of ControlNeXt and Diffusion
$\{Q_t^X, K_t^X, V_t^X\} \leftarrow \epsilon_\theta(z_t^X, c, t, S^X)$
$\{Q_t^Y, K_t^Y, V_t^Y\} \leftarrow \epsilon_\theta(z_t^Y, c, t, S^Y)$
Compute augmented self-attention
$K_t^{\{Y,X\}} \leftarrow \text{Concat}(K_t^Y, K_t^X)$
$A_t^{\{Y,X\}} \leftarrow \frac{Q_t^Y (K_t^{\{Y,X\}})^\top}{\sqrt{d}}$
$A_t^{Y\to Y}, A_t^{Y\to X} \leftarrow A_t^{\{Y,X\}}$
Compute matching via ϕ
$C^{Y\to X} \leftarrow \text{CategoricalMatching}(S^X, S^Y)$
$R_t^{Y\to X} \leftarrow \text{Concat}(A^{Y\to X}, \text{DownSample}(C^{Y\to X}))$
$O^{Y\to X} \leftarrow \phi(R_t^{Y\to X}) + A_t^{Y\to X}$
Predict noise and compute loss
$\hat{\epsilon} \leftarrow \epsilon_\theta(z_t^Y, c, t, O^{Y\to X})$
$\mathcal{L} \leftarrow \|\hat{\epsilon} - \epsilon\|^2$
Update ϕ to minimize $\mathcal{L}$
return ϕ

Algorithm 2 Inference

Input: Exemplar initial latent noise map z_T^X, exemplar and target segmentation maps S^X, S^Y, retrieval flag f
Output: Exemplar and target latent maps z_0^X, z_0^Y

if $f = 1$ **then** ▷ retrieval is enabled
    $S^X \leftarrow$ Automatic Exemplar Retrieval(S^Y)
end if
$z_T^Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, T-1, ..., 1$ **do**
    $\epsilon_\theta^X, \{Q_t^X, K_t^X, V_t^X\} \leftarrow \epsilon_\theta(z_t^X, c, t, S^X)$
    $z_{t-1}^X \leftarrow \text{Sample}(z_t^X, \epsilon_\theta^X)$
    $\epsilon_\theta^Y, \{Q_t^Y, K_t^Y, V_t^Y\} \leftarrow \epsilon_\theta(z_t^Y, c, t, S^Y)$
    $R_t^{Y\to X} \leftarrow \text{Concat}(A_t^{Y\to X}, \text{DownSample}(C^{Y\to X}))$
    $O_t^{Y\to X} \leftarrow \phi(R_t^{Y\to X}) + A_t^{Y\to X}$
    $\tilde{\epsilon} = \epsilon_\theta^Y(z_t^Y, c, t, O_t^{Y\to X})$
    $z_{t-1}^Y \leftarrow \text{Sample}(z_t^Y, \tilde{\epsilon})$
end for
return z_0^X, z_0^Y

Figure 8. **Additional Qualitative Results of AM-Adapter.** Visualization of results generated by **AM-Adapter (Ours)** across various scenarios.

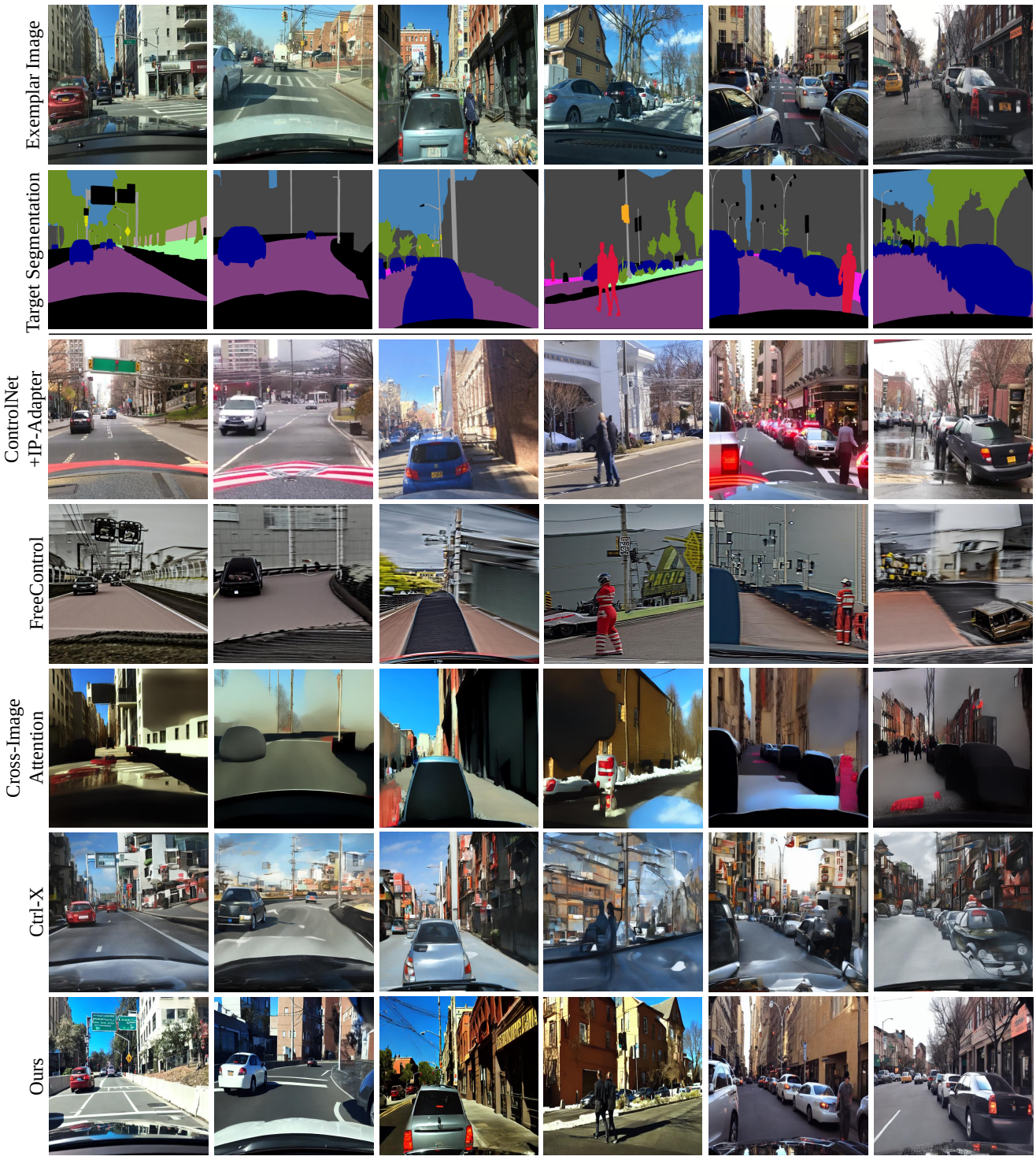


Figure 9. **Additional Qualitative Comparison on BDD100K [32] Dataset.**

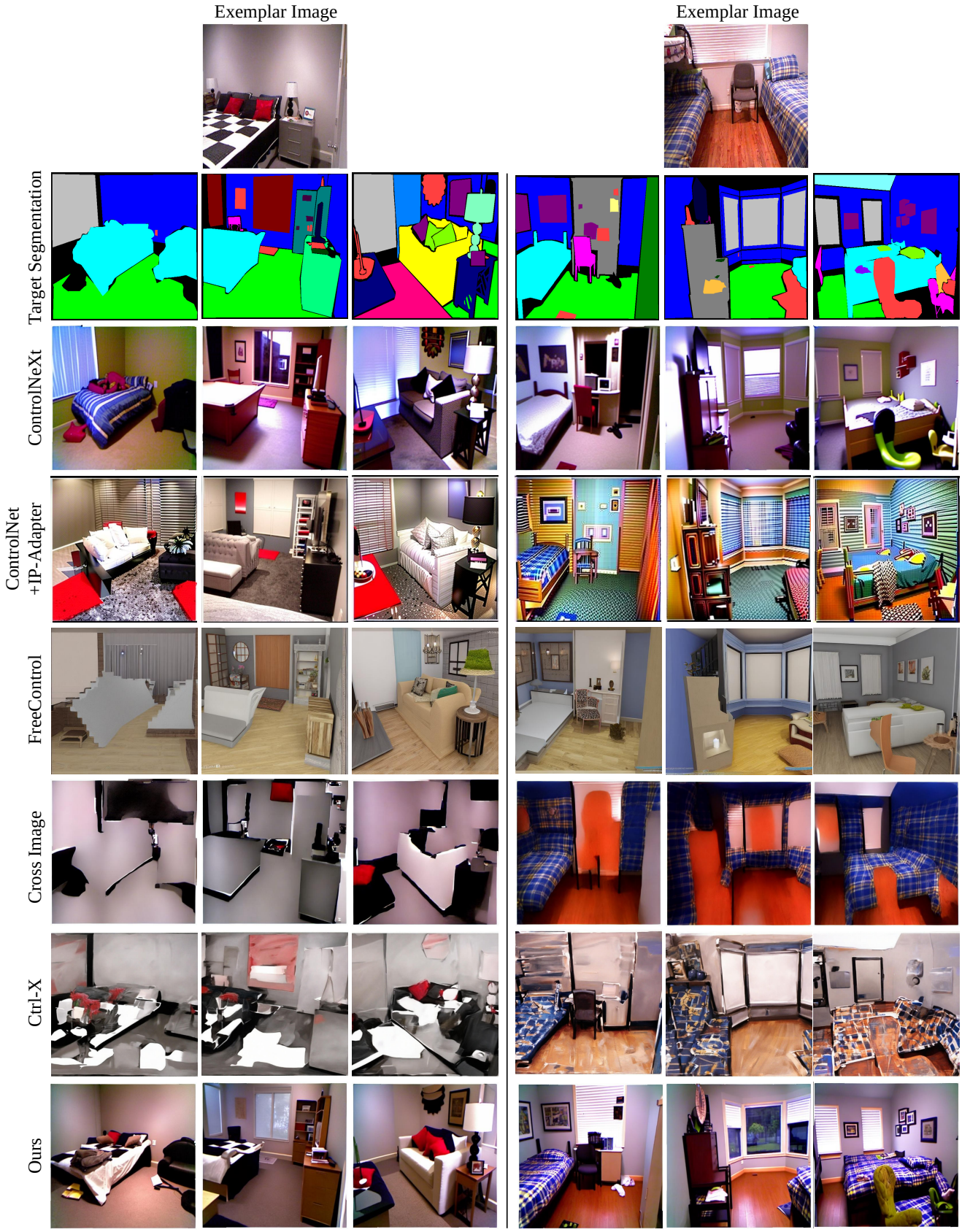


Figure 10. **Additional Qualitative Comparison on NYUv2 [20] Dataset.**

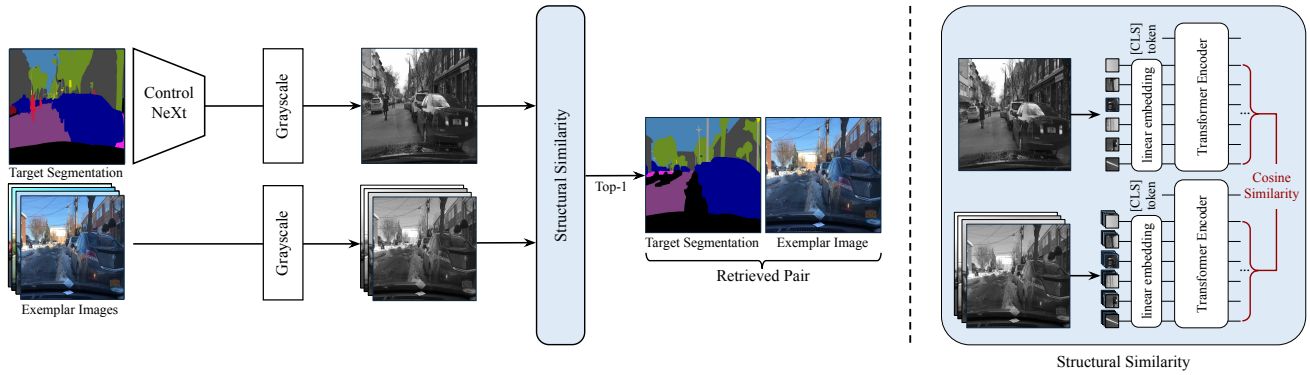


Figure 11. **Details of the Retrieval-based Inference.**

Figure 12. **Retrieval Examples.** During inference, our retrieval technique selects the exemplar image that exhibits the highest structural similarity to the target segmentation map.

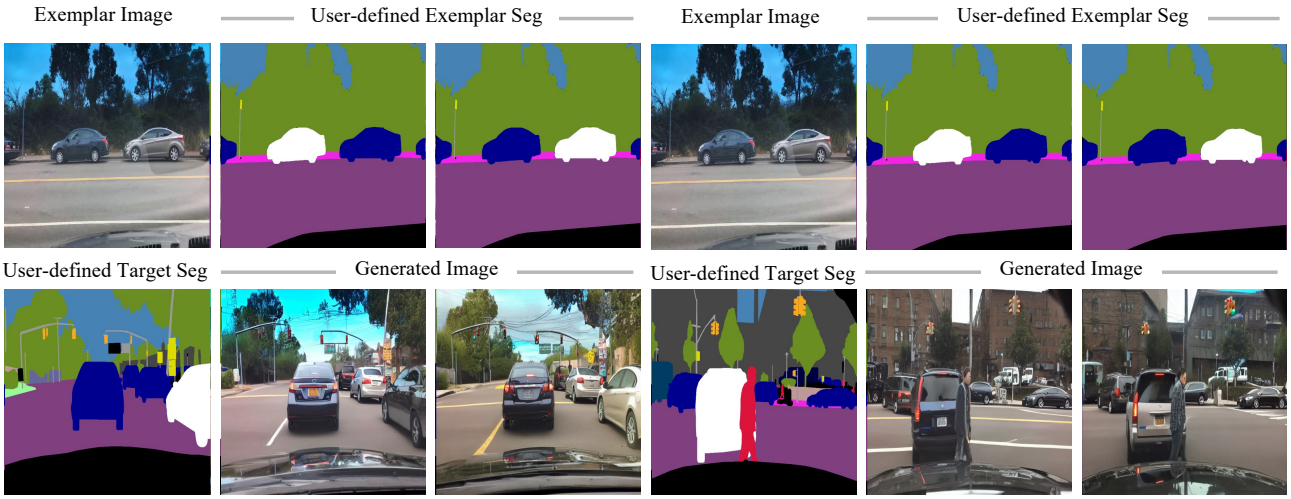


Figure 13. **Application: Controllable One-to-One Matching with User Guidance.** The AM-Adapter enforces one-to-one mapping in a many-to-many setting, enabling controlled appearance transfer. The white regions in the exemplar and target segmentations indicate the source and destination objects, respectively, ensuring accurate appearance mapping.

(a) (b) (c) (d) (e) (f)

Figure 14. **Application: Object Removal by Segmentation-Based Image Editing.** (a) Original target segmentation with desired structure, (b) generated image given the target segmentation map (a), (c) edited target segmentation map, (d) generated image given the target segmentation map (c), (e) edited target segmentation map, (f) generated image given the target segmentation map (e).

(a) (b) (c) (d) (e) (f)

Figure 15. **Application: Object Additon by Segmentation-Based Image Editing.** (a) Edited target segmentation with desired structure, (b) generated image given the target segmentation map (a), (c) edited target segmentation map, (d) generated image given the target segmentation map (c), (e) edited target segmentation map, (f) generated image given the target segmentation map (e).

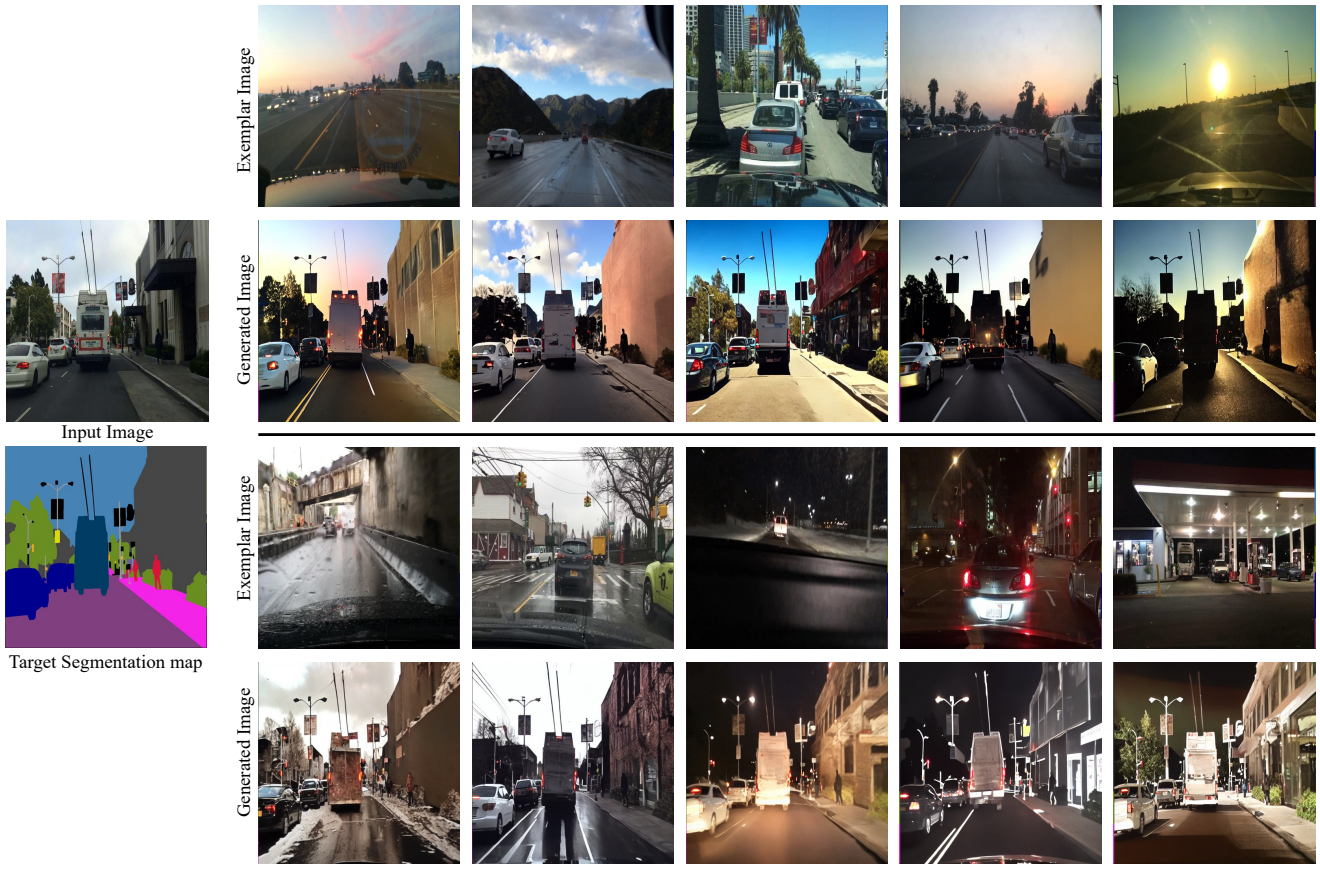


Figure 16. **Application: Image-to-Image Translation.** The first and third row represent exemplar images reflecting various weather conditions and times of day, including various categories such as cloudy days, sunset hours, sunny, night, and rainy conditions. The second and fourth rows depict the resulting images that incorporate the structure of the target segmentation along with each exemplar image.

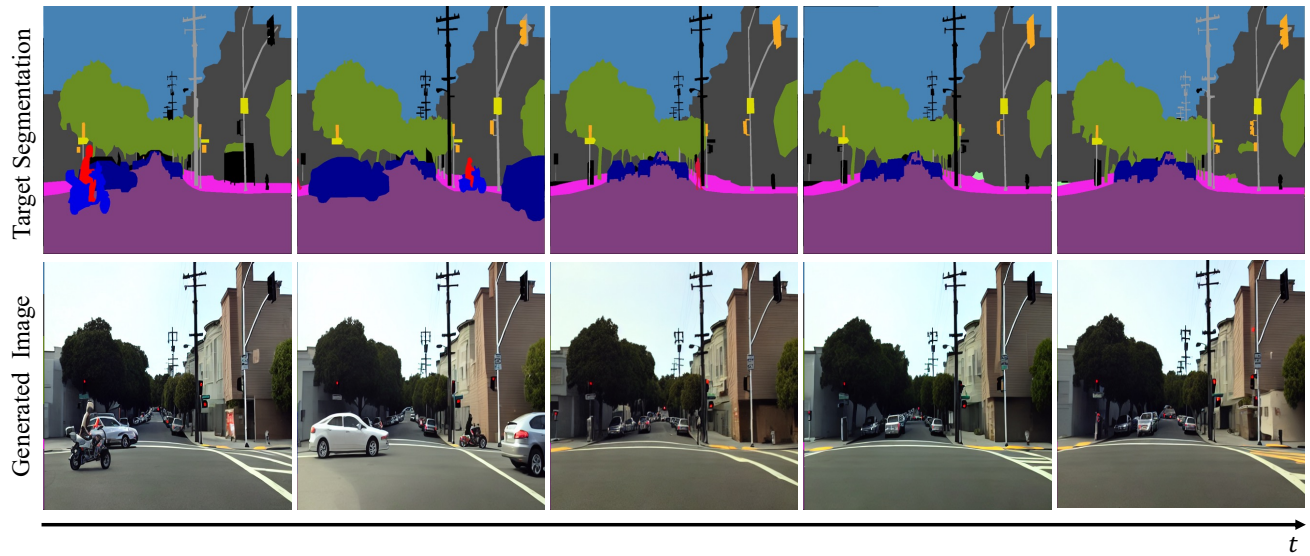


Figure 17. **Application: Appearance-Consistent Consecutive Video Frame Generation.** The target segmentation maps in the first row are consecutive frames provided by the BDD100K [32] dataset. The second row displays the generated image results corresponding to each target segmentation maps. The frames are arranged sequentially from left to right.

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2, 8

[2] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 2, 3, 4, 8

[3] Dimitrios Christodoulou and Mads Kuhlmann-Jørgensen. Finding the subjective truth: Collecting 2 million votes for comprehensive gen-ai model evaluation. *arXiv preprint arXiv:2409.11904*, 2024. 1

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[5] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 1

[6] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 1

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 4

[8] Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *arXiv preprint arXiv:2406.04485*, 2024. 1

[9] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. 1

[10] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pages 129–147. Springer, 2025. 1

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 1

[12] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 1

[13] Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. *arXiv preprint arXiv:2406.07540*, 2024. 1, 2, 8

[14] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2025. 1

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 1

[16] Wuyang Luo, Su Yang, Xinjian Zhang, and Weishan Zhang. Siedob: Semantic image editing by disentangling object and background, 2023. 6

[17] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. 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*, pages 7465–7475, 2024. 2, 8

[18] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 1

[19] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8110, 2024. 1, 2, 3, 4, 8

[20] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 5, 12

[21] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 1

[22] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models, 2024. 1

[23] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 1, 3, 6

[24] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27080–27090, 2024. 1

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[27] Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, Mengping Yang, Cheng Zhang, and Hao Li. Evalalign: Evaluating text-to-image models through precision alignment of multimodal large models with supervised fine-tuning to human annotations. *arXiv preprint arXiv:2406.16562*, 2024. 1

[28] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer, 2022. 1, 4

[29] 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. 1

[30] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20, 2024. 1

[31] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 8

[32] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1, 6, 11, 15

[33] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 8

[34] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 1