UNITY: ATTENTION FLOW NETWORKS FOR ADAPTIVE CONDITIONING IN DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce **UNITY**, Universal-to-Specialized adapter for efficient and scalable composite conditioning in diffusion-based image generation. Unlike prior methods that train separate adapters for each condition, UNITY jointly learns shared semantics across multiple conditioning modalities and later specializes. The proposed two-stage training paradigm consists of a *Universal Stage*, which captures cross-modal joint representations across all conditioning types using half of the total training steps, followed by a *Specialization Stage* that refines modality-specific features with the remaining steps. The core of our design are the **Morphable Attention Flow (MAF) Network** and **Morph Wrapper** modules, which enable channel-aware and spatially adaptive feature alignment through learnable flow fields and attention fusion. This constant-complexity formulation supports flexible operation under single or composite conditioning while significantly reducing inference latency and memory footprint. Extensive experiments on multiple datasets demonstrate that UNITY achieves State-of-the-Art (SoTA) image fidelity and memory efficiency.

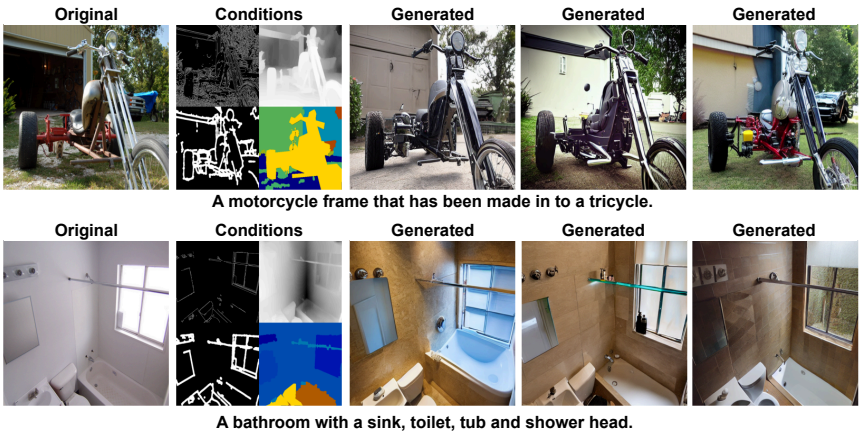


Figure 1: Composite conditioned image generation with UNITY-Adapter. Each example is jointly conditioned on four complementary control signals shown in the **Conditions** column: *top left*, Canny edge; *top right*, Depth map; *bottom left*, Sketch; and *bottom right*, Semantic Segmentation.

1 INTRODUCTION

Conditional image generation has become a foundational capability in modern visual AI, yet as real-world applications increasingly demand simultaneous control over structure, semantics, and style, the scalability of existing conditioning frameworks has emerged as a critical and unsolved bottleneck. Diffusion-based models (Cao et al., 2024; Rombach et al., 2022a;b) have enabled high-fidelity synthesis guided by diverse signals such as text, depth, sketches, and semantic masks, and attention-based

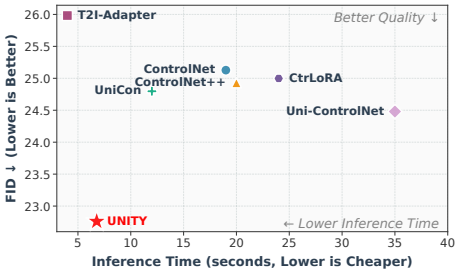


Figure 2: Comparison of FIDs and Infer time across methods.

We are happy to release the code upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

architectures with vision transformers have further strengthened cross-modal alignment (Dosovitskiy et al., 2020; Li et al., 2022; Radford et al., 2021; Karim & Verma, 2026; Das et al., 2026a). Multi-modal conditioning is not a luxury; it is a necessity for structure-preserving tasks such as text-to-image generation (Saharia et al., 2022; Ramesh et al., 2021; 2022) and image-to-image translation (Park et al., 2019; Wang et al., 2018), where spatial layout and semantic consistency are non-negotiable. Composite conditioning, where multiple inputs jointly describe a shared underlying scene, is the natural evolution of this paradigm, offering richer, more coherent guidance than any single modality can provide. Yet the field has treated this as an afterthought, bolting together independent adapters rather than designing systems that fundamentally understand multi-modal structure. This raises an unavoidable question: ***Can we design a single adapter that jointly learns shared semantics across all conditioning modalities, scales to any number of conditions without additional cost, and maintains precise spatial alignment?***

Dominant adapter frameworks, including ControlNet (Zhang et al., 2023), T2I-Adapter (Mou et al., 2024), ControlNet++ (Li et al., 2024a), and CtrLoRA (Xu et al., 2024), treat each conditioning modality as an isolated problem, each requiring a fully independent training cycle to reach optimal performance. The consequence is structural and severe: training cost and memory scale linearly with the number of conditions, making four-modality conditioning demand four times the resources of a single-modality baseline. Uni-ControlNet (Zhao et al., 2023), the closest attempt at unification, still requires $8\times$ the GPU memory of a single-condition setup, with 1271.42M parameters and no mechanism for shared semantic learning across conditions. More fundamentally, none of these architectures possesses a unified mechanism to model fine-grained spatial correspondences across modalities. ControlNet-style adapters duplicate the entire Stable Diffusion encoder, an expensive design choice that paradoxically sacrifices modality-specific adaptability for scale, and ControlNet-style methods execute a parallel copy of the UNet encoder at every denoising step alongside the diffusion backbone, which is the primary driver of inference latency.

We argue that this demands not incremental improvement but a ground-up rethinking of how adapters should handle multiple conditioning modalities. Two core problems must be solved simultaneously, and prior work has ignored both. **First**, efficient multi-modal conditioning requires a framework that jointly learns the shared semantic structure underlying all modalities, rather than treating each in isolation, and does so without any increase in parameter count or training budget as modalities grow. **Second**, spatial correspondence across conditioning modalities cannot be captured by attention alone; it demands explicit, learnable geometric alignment that adapts to each modality's spatial characteristics. We introduce **UNITY-Adapter**, a Universal-to-Specialized framework that confronts both problems directly. UNITY-Adapter trains in two decisive stages: a *Universal Stage* that jointly learns from all K conditions in the first $N/2$ steps, forging a shared discriminative representation space, followed by a *Specialization Stage* that refines each modality individually in the remaining $N/2$ steps, delivering a provable 37.5% reduction in training cost for four conditions with *strictly fixed* parameters regardless of K. Complementing this, the proposed **Morphable Attention Flow (MAF)** network introduces learnable displacement fields that directly capture spatial correspondences between conditioning inputs and latent representations, enabling precise geometric alignment without backbone duplication and without any parallel denoising path, substantially cutting inference latency. The result is striking: UNITY-Adapter achieves SoTA FID and CLIP scores across all four conditioning modalities on a single 24GB GPU, outperforming methods that require $4\times$ to $8\times$ more memory, as shown in Figure 2. Our contributions are summarized as follows:

- We propose **UNITY-Adapter**, a universal-to-specialized framework that enables efficient and flexible composite conditioning without added model complexity or cost.
- We introduce the **Morphable Attention Flow (MAF)** network with **Morph Wrapper** modules, which performs channel-aware and spatially adaptive feature alignment using learnable flow fields and attention-based fusion without duplicating the diffusion backbone.

2 Related Work

Text-to-image diffusion models enable high-fidelity synthesis (Cao et al., 2024; Rombach et al., 2022b;a; Yu et al., 2025; Saharia et al., 2022). ControlNet (Zhang et al., 2023), T2I-Adapter (Mou et al., 2024), Uni-ControlNet (Zhao et al., 2023), UniCon (Li et al., 2024b), ControlNet++ (Li et al., 2024a), UniControl (Qin et al., 2023a), and CtrLoRA (Xu et al., 2024) progressively improve controllable generation. Multi-condition methods (Das et al., 2026b), including Composer (Huang

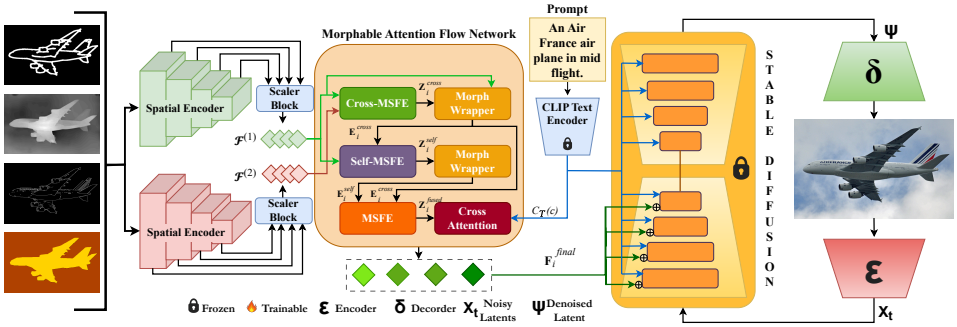


Figure 3: Proposed Framework: Conditioning inputs are processed through dual Spatial Encoders with Scaler Block to produce multi-scale feature pyramids $\mathcal{F}^{(1)}$ and $\mathcal{F}^{(2)}$. The Morphable Attention Flow (MAF) Network performs Cross-Modal alignment (Cross-MSFE + Morph Wrapper), Self-Refinement (Self-MSFE + Morph Wrapper), and Fusion (MSFE), with final features $\mathbf{F}_i^{final}$ conditioned via Cross-Attention with Text Embeddings $\mathcal{C}_T(c)$ and injected into the frozen Stable Diffusion's UNet and fused with Point-Wise Addition.

et al., 2023), multi-modal GANs (Huang et al., 2021), UniCombine (Wang et al., 2025), OminiControl (Tan et al., 2025), and TCIG (Mohamed, 2024), support composite guidance but often require condition-specific training, parallel denoising, composite retraining, or sequential refinement. In contrast, UNITY-Adapter learns shared representations before specialization, enabling flexible composite control with constant complexity and no condition-specific retraining or inference branches.

Flow-based methods provide effective spatial alignment. Normalizing flows (Dinh et al., 2014) introduced invertible transformations, while Diffusion Normalizing Flow (Zhang & Chen, 2021) combines stochastic diffusion (Song et al., 2020) with learnable dynamics. Global-Flow Local-Attention (Ren et al., 2020), Latent Flow Diffusion Models (Ni et al., 2023), and Flow-based Feature Warping (Wei et al., 2020) further exploit flow-guided alignment. Parameter-efficient adapters (Guo et al., 2021; Mei et al., 2026) reduce adaptation cost but remain limited for composite control, while Diff2Flow (Schusterbauer et al., 2025) aligns diffusion and flow matching. Our MAF performs channel-aware attention-driven warping, while UNITY-Adapter couples universal shared training with condition specialization, preserving semantic-spatial consistency and constant training complexity as conditions increase.

3 Methodology

3.1 Preliminaries

Diffusion models generate samples by reversing a gradual noising process. Given a clean image $\mathbf{X}_0$, the forward diffusion process constructs progressively noisier states according to $q(\mathbf{X}_t \mid \mathbf{X}_{t-1}) = \mathcal{N}(\mathbf{X}_t; \sqrt{1-\beta_t}\,\mathbf{X}_{t-1}, \beta_t\mathbf{I})$, where $\{\beta_t\}_{t=1}^T$ is a predefined noise schedule that produces approximately Gaussian noise after T steps. The reverse process aims to denoise by modeling $p_\theta(\mathbf{X}_{t-1} \mid \mathbf{X}_t) = \mathcal{N}(\mathbf{X}_{t-1}; \mu_\theta(\mathbf{X}_t, t), \Sigma_\theta(t))$, typically reparameterized using a noise prediction network ϵ_θ trained via the objective $\mathcal{L}(\theta) = \mathbb{E}_{t,\mathbf{X}_0,\epsilon}[\|\epsilon - \epsilon_\theta(\mathbf{X}_t, t)\|^2]$. In this work, we employ a pretrained Latent Diffusion Model (LDM) based on the Stable Diffusion (Rombach et al., 2022a) architecture, which operates in a compressed latent space using a variational autoencoder (VAE) consisting of an encoder ε, a decoder δ, and a denoising UNet. The encoder maps an input image $\mathbf{X}_0$ to its latent representation $\Psi = \varepsilon(\mathbf{X}_0)$, while the decoder reconstructs the image from the latent as $\hat{\mathbf{X}} = \delta(\Psi)$. Forward diffusion is then applied in the latent space, perturbing Ψ over T timesteps to produce noisy latents $\mathbf{X}_t$. The UNet, parameterized by θ, performs reverse denoising according to $\mathbf{X}_{t-1} = f_\theta(\mathbf{X}_t, t, c)$, where c denotes the text-conditioning vector obtained from a frozen CLIP text encoder $\mathcal{C}_T$. The conditioning is integrated into the UNet via cross-attention at each step.

3.2 Framework

UNITY operates as a Plug-and-Play module for the frozen SD1.5 UNet backbone and text-encoder $\mathcal{C}_\mathcal{T}$. The core innovation is our Flow-based Adapter design (Fig. 3) coupled with the two-stage training paradigm in Algorithm 1 that yields $\frac{K-1}{2K}$ training reduction (37.5% for $K = 4$). Rather than parallel condition-specific denoising, UNITY-Adapter provides global multi-scale guidance to the frozen SD1.5, which performs core denoising while the adapter ensures conditional alignment across

timesteps. The adapter maintains constant parameters regardless of K, enabling flexible single or composite conditioning at inference. Text embeddings $\mathcal{C}_\mathcal{T}$ maintain semantic coherence between visual conditions and textual intent, enhancing structural consistency without compromising generative quality or computational efficiency.

3.3 Adapter Architecture

UNITY introduces a unified conditioning mechanism for the frozen Stable Diffusion UNet through two key components: *Spatial Encodings* and the *Morphable Attention Flow (MAF) Network*, which consists of *Multi-Scale Flow Estimators* (MSFE) and *Morph Wrappers*. The spatial encoders transform conditioning inputs into hierarchical feature pyramids that are modulated by a Scaler Block and aligned with the UNet's multi-scale architecture. The MAF Network then processes these features through three sequential stages: *Cross-Modal Alignment*, where a Cross-MSFE followed by a Morph Wrapper establishes geometric correspondences between complementary conditioning features; *Self-Refinement*, where a Self-MSFE and Morph Wrapper further enhance intra-modal structural consistency; and *Fusion*, where a refinement MSFE integrates the cross-aligned and self-refined representations into unified multi-scale embeddings. Finally, the fused features are conditioned through Cross-Attention with the text embeddings $\mathcal{C}_T(c)$ before being injected into the frozen Stable Diffusion UNet via point-wise addition.

3.3.1 Spatial Encoding

The spatial encoding module transforms conditional inputs into hierarchical feature representations aligned with the UNet's multi-scale architecture. It comprises two primary components: a *Spatial Encoder*, which extracts multi-scale features, and a *Scaler Block*, which adaptively modulates these features before integration into the UNet.

Spatial Encoder: The spatial encoder constructs a hierarchical feature pyramid that preserves spatial fidelity while aligning with the multi-scale UNet structure. Pixel unshuffle performs a lossless spatial-to-channel transformation, and pre-activation residual blocks ensure stable gradient propagation. Given K conditioning inputs $\mathbf{C} \in \mathbb{R}^{H\times W\times C_k}$ with $C_k = 3K$[1], we first apply pixel unshuffle with scale factor s:

$$\mathbf{F}_0 = \mathit{PixelUnshuffle}(\mathbf{C}), \quad \mathbf{F}_0 \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times C_k s^2}. \tag{1}$$

The encoder then produces multi-level features $\{\mathbf{F}_i\}_{i=1}^{L}$ through L hierarchical stages:

$$\mathbf{F}_i = \mathcal{R}_i^{(2)}\big(\mathcal{R}_i^{(1)}(\mathcal{D}_i(\mathbf{F}_{i-1}))\big), \quad i = 1,\dots,L. \tag{2}$$

Here, $\mathcal{D}_i$ denotes a strided 3×3 convolution that reduces spatial resolution and increases channel dimension from d_{i-1} to d_i. Each $\mathcal{R}_i^{(j)}$ is a pre-activation residual block defined as, $\mathcal{R}(\mathbf{X}) = \mathbf{X}+\mathcal{G}(\mathbf{X})$, where $\mathcal{G}(\cdot)$ consists of two 3×3 convolutions with BN and ReLU in pre-activation order. All convolutions use padding to preserve spatial resolution, and bias terms are omitted due to normalization. This formulation yields a computationally efficient multi-scale encoder while retaining dense spatial information.

Scaler Block: The Scaler Block applies adaptive channel-wise modulation to encoded features before fusion. This adaptive modulation enables coherent feature fusion across heterogeneous conditioning types while preserving spatial fidelity. For each encoder output $\mathbf{F}_i \in \mathbb{R}^{H_i\times W_i\times d_i}$, the block performs a learnable 1×1 transformation:

$$\tilde{\mathbf{F}}_i = \mathit{Conv}_{d_i\to d_i}^{1\times1}(\mathbf{F}_i), \quad i \in \{1,\dots,L\}. \tag{3}$$

These pointwise convolutions act as gating mechanisms that emphasize or suppress channels based on the conditioning modality, enabling dynamic feature reweighting while maintaining spatial structure. The set of modulated features $\{\tilde{\mathbf{F}}_i\}_{i=1}^{L}$ is further processed through the MAF Network.

3.3.2 Morphable Attention Flow Network

The *Morphable Attention Flow (MAF) Network* constitutes the core processing that transforms spatial embeddings through iterative refinement via flow estimation and morphological wrapping. Operating on feature pyramids from dual spatial encoders, the MAF Network performs cross-modal

[1]During Stage 1 (Universal Training), all K RGB conditioning modalities are jointly provided as channel-wise inputs, yielding $C_k = 3K$. During Stage 2 (Specialization), the model is trained with an individual conditioning modality, corresponding to $C_k = 3$.

Algorithm 1: UNITY: Two-Stage Training

Input: $\{C_i\}_{i=1}^{K}$, N, frozen UNet θ
Universal Stage ($N/2$ **steps**)
$\phi \leftarrow \text{Pretrain}(\{C_i\}, N/2)$
Specialization Stage ($N/2$ **steps**)
for $i = 1$ to K **do**
    $\phi_i \leftarrow \text{Finetune}(C_i, N/2)$
end for
$\phi \leftarrow \bigcup_i \phi_i$
$T_{UNITY} = \frac{N}{2}(1 + K)$ vs. KN

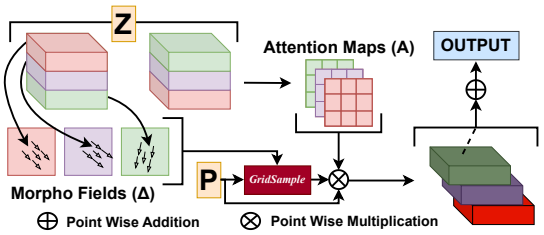


Figure 4: Morph Wrapper: Adaptive feature warping using learned Morpho Fields $\mathbf{\Delta}$ and Attention Weighted Aggregation across M sampling points.

alignment and self-refinement across L hierarchical scales. Two parallel spatial encoding pathways process complementary conditioning signals, each producing multi-scale feature pyramids through the spatial encoding module described previously. The outputs from the first and second spatial encoders, after their respective *Scaler Blocks*, yield two feature sets $\mathcal{F}^{(1)} = \{\tilde{\mathbf{F}}_i^{(1)}\}_{i=1}^{L}$ and $\mathcal{F}^{(2)} = \{\tilde{\mathbf{F}}_i^{(2)}\}_{i=1}^{L}$, where $\tilde{\mathbf{F}}_i^{(1)}, \tilde{\mathbf{F}}_i^{(2)} \in \mathbb{R}^{C_i \times H_i \times W_i}$ represent modulated features at scale i. These dual representations capture complementary structural and semantic information that will be progressively refined through the MAF Network's stage processing design. For each scale $i \in \{1, \ldots, L\}$, processing begins with cross-modal flow estimation, where features from both encoders are concatenated and fed into a MSFE (Section 3.3.3) to establish inter-modal correspondences:

$$\mathbf{Z}_i^{cross} = \mathit{MSFE}_{cross}\Big(\text{Concat}\big(\tilde{\mathbf{F}}_i^{(1)}, \tilde{\mathbf{F}}_i^{(2)}\big)\Big), \tag{4}$$

where $\mathbf{Z}_i^{cross}$ encodes cross-modal flow features. This $MSFE_{cross}$ output captures the geometric and semantic relationships between the two encoding features, providing the foundation for subsequent morphological alignment. The cross-modal features $\mathbf{Z}_i^{cross}$ are then processed by a Morph Wrapper (Section 3.3.4), which warps the primary encoder features $\tilde{\mathbf{F}}_i^{(1)}$ using displacement fields and attention maps derived from the cross-modal flow:

$$\mathbf{E}_i^{cross} = \mathit{MorphWrapper}\big(\tilde{\mathbf{F}}_i^{(1)}, \mathbf{Z}_i^{cross}\big), \tag{5}$$

where $\mathbf{E}_i^{cross}$ represents the cross-aligned embeddings. The *cross Morph Wrapper* extracts displacement fields $\mathbf{\Delta}_i^{cross}$ and attention maps $\mathbf{A}_i^{cross}$ from $\mathbf{Z}_i^{cross}$, performing adaptive spatial warping that aligns multi-modal features through learned geometric transformations. Following cross-modal alignment, the network performs self-refinement to enhance intra-modal consistency. A $MSFE_{self}$ processes the concatenation of the original primary features and the cross-aligned embeddings:

$$\mathbf{Z}_i^{self} = \mathit{MSFE}_{self}\Big(\text{Concat}\big(\tilde{\mathbf{F}}_i^{(1)}, \mathbf{E}_i^{cross}\big)\Big), \tag{6}$$

where $\mathbf{Z}_i^{self}$ captures self-consistent flow patterns that refine the representation while preserving structure. Unlike the cross-modal stage, the self Morph Wrapper operates exclusively on the self-refined features, using $\mathbf{Z}_i^{self}$ as both the embedding source and the flow guidance:

$$\mathbf{E}_i^{self} = \mathit{MorphWrapper}\big(\mathbf{Z}_i^{self}, \mathbf{Z}_i^{self}\big), \tag{7}$$

where $\mathbf{E}_i^{self}$ represents the self-aligned embeddings. The *self Morph Wrapper* similarly extracts displacement fields $\mathbf{\Delta}_i^{self}$ and attention maps $\mathbf{A}_i^{self}$ from $\mathbf{Z}_i^{self}$, performing adaptive warping that preserves structural integrity while refining local spatial details. This self-referential deformation enables fine-grained correction and enhancement without interference from cross-modal signals. The outputs from both cross-modal and self-refinement pathways are then fused through a refinement MSFE that integrates complementary information:

$$\mathbf{Z}_i^{fused} = \mathit{MSFE}_{refine}\Big(\mathit{Concat}\big(\mathbf{E}_i^{cross}, \mathbf{E}_i^{self}\big)\Big), \tag{8}$$

which produces unified flow-estimated embeddings $\mathbf{Z}_i^{fused}$ that combine cross-modal alignment with self-consistent refinement. These fused features capture both inter-modal correspondences and intra-modal structural coherence. Unlike the T2I-Adapter, to ensure semantic guidance to the

adapter with the text prompt, the fused embeddings are conditioned through cross-attention with text embeddings from the frozen CLIP text encoder $\mathcal{C}_T$:

$$\mathbf{F}_i^{final} = \textit{CrossAttention}\big(\mathbf{Z}_i^{fused}, \mathcal{C}_T(c)\big), \tag{9}$$

where c denotes the text prompt and $\mathbf{F}_i^{final}$ represents the semantically aligned feature embedding at scale i. This cross-attention mechanism ensures that the visual conditioning remains consistent with the textual intent throughout the generation process.

3.3.3 Multi-Scale Flow Estimator (MSFE)

Multi-Scale Flow Estimator (MSFE) serves to learn hierarchical feature transformations that estimate the Morpho Fields and attention weights for flow-aware guidance across spatial scales. It consists of a lightweight convolutional hierarchy that progressively refines latent representations through multi-resolution filtering and non-linear activation to produce morphological displacement fields and their corresponding attention maps. This lightweight architecture enables rapid convergence during two-stage training while maintaining sufficient capacity for multi-scale flow patterns. Given an input feature map $\mathbf{Z}_{in} \in \mathbb{R}^{C_z \times H_w \times W_w}$, the MSFE produces an output $\mathbf{Z}_{out} \in \mathbb{R}^{(3M) \times H_w \times W_w}$ through a sequence of convolutional and activation layers:

$$\begin{aligned} \mathbf{Z}_1 &= LReLU\big(Conv^{3\times3}_{C_z \to C_1}(\mathbf{Z}_{in})\big), \quad \mathbf{Z}_2 = LReLU\big(Conv^{3\times3}_{C_1 \to C_2}(\mathbf{Z}_1)\big), \\ \mathbf{Z}_3 &= LReLU\big(Conv^{3\times3}_{C_2 \to C_3}(\mathbf{Z}_2)\big), \quad \mathbf{Z}_{out} = Conv^{3\times3}_{C_3 \to 3M}(\mathbf{Z}_3), \end{aligned} \tag{10}$$

where $LReLU$ denotes the Leaky ReLU activation with a negative slope of 0.1, and M is the number of sampling points.

The output $\mathbf{Z}_{out}$ is then decomposed through channel splitting to extract the Morpho Fields and attention weights:

$$\mathbf{\Delta} = \mathbf{Z}_{out}[0:2M, :, :], \quad \mathbf{A} = \mathbf{Z}_{out}[2M:3M, :, :], \tag{11}$$

where the first $2M$ channels encode the Morpho Fields $\mathbf{\Delta} \in \mathbb{R}^{M \times 2 \times H_w \times W_w}$ defining morphological spatial offsets (with 2 coordinates per sampling point), and the remaining M channels represent the morphological attention weights $\mathbf{A} \in \mathbb{R}^{M \times H_w \times W_w}$. The combined representation $[\mathbf{\Delta}, \mathbf{A}] = \mathbf{Z}_{out}$ directly serves as the conditioning feature for the *MorphWrapper* function, enabling end-to-end learning of morphological transformations.

3.3.4 Morph Wrapper

As illustrated in Fig. 4, the *Morph Wrapper* operates as a morphological attention mechanism that unifies spatial embeddings and attention maps through adaptive sampling. Unlike conventional deformable attention with fixed offset patterns, the Morph Wrapper dynamically wraps features, aligning multi-scale embeddings to context-dependent spatial configurations via learned attention modulation to preserve spatial information that is typically lost in standard cross-attention. The *MorphWrapper* function takes two inputs: a source embedding $\mathbf{P} \in \mathbb{R}^{C_p \times H_w \times W_w}$ and the MSFE output $\mathbf{Z} \in \mathbb{R}^{(3M) \times H_w \times W_w}$, where M denotes the number of sampling points. $\mathbf{Z}$ contains both Morpho Fields and attention maps $[\mathbf{\Delta}, \mathbf{A}] = \mathbf{Z}$, with $\mathbf{\Delta} \in \mathbb{R}^{M \times 2 \times H_w \times W_w}$ representing morphological spatial offsets that define adaptive sampling locations, and $\mathbf{A} \in \mathbb{R}^{M \times H_w \times W_w}$ denoting attention weights. The mechanism synthesizes wrapped features via differentiable bilinear interpolation:

$$\mathbf{P}'_m = \textit{GridSample}\big(\mathbf{P}, \mathbf{\Delta}_m\big), \quad m = 1, \ldots, M. \tag{12}$$

Each warped embedding is modulated by its attention map through softmax-normalized aggregation:

$$\textit{MorphWrapper}(\mathbf{P}, \mathbf{Z})(p) = \sum_{m=1}^{M} \frac{\exp(\mathbf{A}_m(p))}{\sum_{j=1}^{M} \exp(\mathbf{A}_j(p))} \mathbf{P}'_m(p), \tag{13}$$

where p indexes spatial positions. This differentiable morphological attention mechanism enables flexible geometric adaptation, enhancing correspondence between conditional inputs and latent representations while maintaining end-to-end trainability.

3.4 Fusion with Backbone

At each of the L scales, adapter outputs the final condition-aware features $\{\mathbf{F}_i^{final}\}_{i=1}^{L}$ (Eq. 9). These features are integrated with the corresponding encoder stages of the frozen Stable Diffusion 1.5 UNet through point-wise addition:

$$\mathbf{U}_i^{final} = \mathbf{U}_i^{enc} + \mathbf{F}_i^{final}, \tag{14}$$

Table 1: Comparison of FID (↓) and CLIP (↑) scores across 4 conditions. Model complexity is reported in terms of Parameters (M), FLOPs (G), and Memory (GB), where memory denotes the total training memory required to load SD1.5 with the corresponding adapter. FLOPs are measured with a batch size of 1, while training memory is reported with a batch size of 4. Best are in **bold**.

Model	Params. (M)	FLOPs (G)	Memory (GB)	Canny		Depth		Sketch		Segmentation	
				FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑
ControlNet (Zhang et al., 2023)	361.28	116.61	24×2	22.84	27.41	25.68	27.51	24.93	27.38	27.06	27.04
T2I-Adapter (Mou et al., 2024)	77.37	29.97	24×1	23.73	26.72	26.03	26.46	26.51	27.21	27.66	26.99
ControlNet++ (Li et al., 2024a)	361.28	116.61	24×2	23.59	27.09	25.19	27.56	24.71	27.15	26.23	27.11
Uni-ControlNet (Zhao et al., 2023)	1271.42	210.77	24×8	23.11	27.21	24.92	27.43	24.56	27.54	25.33	27.49
CtrLoRA (Xu et al., 2024)	398.28	135.15	24×4	22.59	27.16	25.65	26.34	26.02	25.46	25.73	25.62
UniCon (Li et al., 2024b)	150.00	111.62	24×8	22.86	26.92	25.30	27.42	24.61	27.13	26.42	27.04
UNITY$_{Ind}$	365.25	135.82	24×1	22.37	28.09	24.12	27.90	24.38	27.76	25.18	27.89
UNITY$_{Pre}$	365.25	135.82	24×1	**21.48**	**28.52**	**22.44**	**28.18**	**23.21**	**28.54**	**23.91**	**27.91**

Table 2: Comparison of FID (↓) and CLIP (↑) scores across 4 conditions. Model complexity includes Parameters (M), and FLOPs (G) for each adapter with the SDXL backbone. FLOPs are measured with a batch size of 1. Best are in **bold**.

Model	Params (M)	FLOPs (G)	Canny		Depth		Sketch		Segmentation	
			FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑
ControlNet (Zhang et al., 2023)	1250.98	1336.69	23.75	31.71	26.42	31.68	25.87	31.64	27.82	31.58
T2I-Adapter (Mou et al., 2024)	79.03	29.95	24.89	31.42	27.18	31.28	27.64	31.36	28.91	31.19
ControlNet++ (Li et al., 2024a)	1250.98	1336.69	24.12	31.58	26.09	31.54	25.43	31.61	27.35	31.52
UNITY$_{Ind}$	384.83	140.17	21.18	31.86	23.42	31.73	23.89	31.92	24.54	31.68
UNITY$_{Pre}$	384.83	140.17	**20.35**	**32.18**	**21.26**	**31.89**	**21.98**	**32.21**	**22.67**	**31.76**

where $\mathbf{U}_i^{enc}$ denotes the intermediate feature map from the UNet encoder at stage i, and $\mathbf{U}_i^{final}$ represents the fused encoder representation after adapter injection. This point-wise additive fusion retains the UNet's pretrained representational strength while infusing globally consistent, structure-aware conditioning from the Adapter. Through staged flow estimation from the *Morpho Fields*, self and cross refinements through *MSFE*, and text-conditioned guidance, the adapter progressively enhances spatial coherence and semantic alignment across scales, ensuring efficient and stable conditioning throughout the denoising process.

4 Experiments

In this section, we present the experimental evaluation of UNITY using both SD1.5 and SDXL backbones. We first describe the datasets and evaluation setup, followed by the main quantitative and qualitative results and model complexity analysis.

Datasets: We train UNITY on MS-COCO 2017 (Lin et al., 2014) and MultiGen-20M (Qin et al., 2023b) datasets. MS-COCO provides diverse real-world images with rich annotations across multiple object categories, while MultiGen-20M offers large-scale multi-modal conditioning pairs essential for learning universal cross-modal representations. For evaluation, we utilize the MS-COCO 2017 validation set comprising 5,000 images with 25,000 image-caption pairs, enabling assessment of generation quality and fidelity.

4.1 Results and Discussion

Table 1 presents a comprehensive comparison of UNITY-Adapter against SoTA conditional image generation methods across four conditioning modalities. We analyze model complexity, quantitative performance, and the effectiveness of our training paradigm.

Quantitative Performance Discussion. As shown in Table 1, UNITY$_{Pre}$ achieves the best FID and CLIP across all four SD1.5 conditions, averaging 22.76 FID and 28.29 CLIP, improving over Uni-ControlNet (24.48, 27.42) by 7.03% and 3.17%, respectively. Gains are largest on Depth, where FID drops from 24.92 to 22.44 (9.95%) and CLIP reaches 28.18, exceeding ControlNet++ (27.56). Improvements on Canny (21.48 FID, 28.52 CLIP) and Sketch (23.21 FID) further demonstrate consistent structural and semantic gains across modalities. Notably, UNITY$_{Pre}$ also uses substantially fewer parameters and FLOPs than Uni-ControlNet, highlighting its efficiency.
Table 2 shows the same trend on SDXL, where UNITY$_{Pre}$ obtains the lowest FID for Canny (20.35), Depth (21.26), Sketch (21.98), and Segmentation (22.67), alongside the highest CLIP scores, reaching 32.21. On Canny, it improves FID over ControlNet by 14.32% (23.75→20.35), with consistent gains elsewhere. Universal pretraining further improves FID over UNITY$_{Ind}$ by 5.22% on SD1.5, with comparable SDXL gains, including 4.91% on Depth. These results show that joint cross-condition pretraining strengthens structural consistency and semantic robustness without additional inference-time overhead.



Figure 5: Qualitative results across conditions shows that UNITY consistently produces the most accurate and semantically aligned outputs.

Qualitative Analysis: Figure 5 compares UNITY against 6 baseline models across 4 diverse prompts, testing attribute fidelity and concept fusion. In *Row 1*, UNITY generates realistic hangar scenes with accurate lighting and spatial composition, while baselines suffer color degradation: UniCon produces oversaturated blue-yellow tones; CtrLoRA introduces heavy yellow-orange casts, ControlNet++ displays unnatural blue tinting; Uni-ControlNet generates excessive yellow coloring; T2I-Adapter creates harsh contrast, destroying detail; and ControlNet exhibits orange-red contamination, eliminating photorealism. *Row 2* demonstrates UNITY's superior concept fusion, integrating clock faces within bicycle wheels with clear visibility, whereas UniCon produces abstract forms, CtrLoRA generates poor detail clarity, ControlNet++ creates blurry outputs, Uni-ControlNet obscures features, T2I-Adapter adds distracting backgrounds, and ControlNet compromises legibility.
Row 3 shows UNITY maintaining photorealistic aircraft quality while baselines produce overexposure (UniCon), invalid coloring (CtrLoRA, Uni-ControlNet, T2I-Adapter), a toy-like appearance (ControlNet++), or pure silhouettes (ControlNet). *Row 4* highlights UNITY's accurate motorcycle customization rendering versus baselines' abstract renderings, poor detail, excessive coloring, or texture destruction. UNITY consistently maintains photorealism, accurate colors, and compositional coherence while baselines systematically fail through color contamination and detail loss.

Model Complexity Analysis: As summarized in Tables 1 and 2, UNITY achieves SoTA FID and CLIP with competitive efficiency. On SD1.5, it uses 365.25M parameters and 135.82G FLOPs, comparable to ControlNet (361.28M, 116.61G), but requires only 24×1 memory versus 24×2–24×4 for ControlNet, ControlNet++, and CtrLoRA, and 24×8 for Uni-ControlNet/UniCon. Despite a $3.48\times$ parameter reduction over Uni-ControlNet (1271.42M, 210.77G), UNITY$_{Pre}$ achieves lower FID across all conditions, demonstrating a stronger performance-efficiency trade-off.

On SDXL, UNITY requires only 384.83M parameters and 140.17G FLOPs versus 1250.98M and 1336.69G for ControlNet and ControlNet++, while improving Canny FID from 23.75 to 20.35 and CLIP from 31.71 to 32.18. Although moderately heavier than T2I-Adapter (79.03M, 29.95G), it achieves substantially better FID, e.g., 21.26 versus 27.18 on Depth. With fixed adapter size across K conditions and shared universal pretraining, UNITY reduces four-condition training cost by 37.5% while retaining top-tier generation quality within a single-condition memory budget.

4.2 Ablation Study

In Table 3 and 4, we systematically validate each component of UNITY-Adapter through targeted ablations spanning text conditioning, architectural design, and multi-condition pretraining. All variants use identical training settings and are evaluated on the MS-COCO validation set with Sketch and Segmentation conditioning, showing that every proposed component contributes meaningfully.

***Case 1* - No-Text-Prompt Settings:** We evaluate two increasingly restrictive settings: removing prompts from the adapter only (*w/o P2A*) and from both Stable Diffusion and the adapter (*w/o P2SD+A*). As shown in Table 3, **UNITY**$_{Pre}$ consistently outperforms all baselines across both structural modalities. Under *w/o P2A*, it reduces FID by 5–7 points over strong competitors such as CtrLoRA and Uni-ControlNet while improving CLIP by $+7$ to $+8$ points, indicating stronger semantic consistency without adapter-level text conditioning. When text is removed entirely (*w/o P2SD+A*),

baselines degrade sharply, with FID increasing by ∼6–8 points and CLIP decreasing by ∼3 points, revealing substantial dependence on textual supervision. In contrast, **UNITY**$_{Pre}$ retains the lowest FID (28.83 / 30.08) and highest CLIP (24.16 / 23.20) for Sketch and Segmentation, respectively. This robustness demonstrates that unified pretraining internalizes strong structural priors, enabling high-quality, semantically coherent, structure-driven generation even without text guidance.

***Case 2* - Architectural Design Variants:** Table 4 validates the proposed flow-based architecture by replacing key MAF components with standard attention mechanisms. *Cross-modal Flow w/ Cross-Attention* worsens Sketch FID by 9.2% (25.35) and CLIP by 4.3% (27.31), showing that explicit spatial modeling is more effective than attention-only alignment. Replacing self-refinement with self-attention further degrades performance (FID 25.88, CLIP 27.13), confirming the benefit of geometrically adaptive morphological warping over global relation modeling. Removing text cross-attention produces the largest CLIP drop of 1.89 points (6.6%), highlighting its role in semantic alignment. Finally, *Single Spatial Encoder* causes the largest FID increase (+3.19, +13.7%), demonstrating that dual-pathway encoding captures richer cross-modal representations than sequential processing.

Table 3: Ablation (Case 1): No-Text-Prompt Settings. *Without Prompt to Adapter (w/o P2A)* and *Without Prompt to Stable Diffusion and Adapter (w/o P2SD+A)*.

Model	w/o P2A				w/o P2SD+A			
	Sketch		Segmentation		Sketch		Segmentation	
	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑
ControlNet (Zhang et al., 2023)	32.21	19.23	31.62	19.30	38.84	16.98	37.72	15.97
T2I-Adapter (Mou et al., 2024)	–	–	–	–	36.59	17.53	35.50	16.45
ControlNet++ (Li et al., 2024a)	32.28	19.26	31.68	19.33	38.90	17.02	37.81	16.01
Uni-ControlNet (Zhao et al., 2023)	30.82	20.16	30.07	20.20	36.28	16.87	35.44	16.47
CtrLoRA (Xu et al., 2024)	30.12	20.20	29.89	20.24	35.34	16.91	34.52	16.50
UniCon (Li et al., 2024b)	30.91	19.76	31.12	19.49	37.10	17.61	35.43	17.97
UNITY$_{Pre}$	**24.94**	**27.58**	**25.80**	**27.04**	**28.83**	**24.16**	**30.08**	**23.20**

Table 4: Comprehensive ablation study of UNITY-Adapter. Three experimental groups analyze: (1) text-conditioning removal, (2) architectural design components, and (3) multi-condition pretraining strategies. FID (↓) and CLIP (↑) scores are reported for Sketch and Segmentation conditions.

Cases	Sketch		Segmentation	
	FID ↓	CLIP ↑	FID ↓	CLIP ↑
***Case 2:* Architectural Design Variants**				
Cross-modal Flow w/ Cross-Attention	25.35	27.31	26.29	26.76
Self-refinement w/ Self-Attention	25.88	27.13	26.75	26.54
w/o Text Cross-Attention	24.73	26.65	25.65	25.98
Single Spatial Encoder	26.40	26.92	27.38	26.33
***Case 3:* Multi-Condition Pretraining**				
Canny + Depth	24.79	27.48	25.63	26.93
Depth + Segmentation	24.64	27.63	25.45	27.07
Depth + Sketch + Segmentation	23.94	28.12	24.77	27.53
Canny + Sketch + Segmentation	23.88	28.17	24.70	27.58
UNITY$_{Pre}$	**23.21**	**28.54**	**23.91**	**27.91**

***Case 3* - Multi-Condition Pretraining Strategies:** Progressive improvement with modality diversity validates our universal pretraining paradigm, as can be observed in Table 4. Two-modality configurations (*Canny + Depth*, *Depth + Segmentation*) average FID 1.57 above baseline and CLIP 0.95 below, establishing dual-condition baselines. Three-modality combinations narrow the gap: *Depth + Sketch + Segmentation* achieves FID 23.94 (+0.73, +3.1%) and CLIP 28.12 (-0.42, -1.5%), while *Canny + Sketch + Segmentation* reaches FID 23.88 and CLIP 28.17. The consistent trend shows that increased modality diversity during universal pretraining establishes stronger shared representations. Our full four-modality pretraining (Canny, Depth, Sketch, Segmentation) achieves optimal performance (FID 23.21, CLIP 28.54), demonstrating that maximal conditioning diversity enables the most robust structure-aware semantic learning. This validates the universal-to-specialized training paradigm's ability to leverage cross-modal information effectively.

5 Conclusion

In this work, we presented UNITY, a novel adapter that fundamentally addresses scalability and efficiency limitations in multi-modal conditional diffusion models. Our two-stage training paradigm enables significant training reduction while maintaining constant parameters regardless of conditioning diversity. The proposed novel Morphable Attention Flow Network with Dual Spatial Encoders, Multi-Scale Flow Estimators, and Morph Wrapper modules provides precise geometric alignment through learned displacement fields and attention-based fusion to preserve spatial information. Unlike existing methods that scale linearly in training cost and require separate adapters per modality, UNITY-Adapter offers global guidance without redundant parallel denoising, substantially reducing computational overhead while supporting flexible single or composite conditioning at deployment. This work presents an efficient paradigm for controllable generation, showing that principled architecture and strategic training can effectively bridge generalization and specialization.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of the manuscript, and for generating illustrative code snippets during development. We have not used generative AI tools for formulating the research ideas, designing the experiments, or interpreting the results. All AI-assisted text has been carefully reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, user studies, or the collection of new personal data. All experiments use publicly available datasets (MS-COCO 2017 and MultiGen-20M) and publicly released pretrained models (Stable Diffusion v1.5 and SDXL), used in accordance with their respective licenses. As a controllable image generation method, UNITY shares the dual-use risks of text-to-image diffusion models, including the potential creation of misleading or harmful imagery, and structure-preserving conditions such as pose or depth may make such content more convincing. UNITY also inherits the social and demographic biases present in its frozen backbone and training data. We encourage responsible deployment, including the use of the safety filters and usage policies that accompany the underlying diffusion models, and will release our code under terms consistent with these licenses.

REPRODUCIBILITY STATEMENT

Appendix A provides the complete training and evaluation protocol, including dataset details, optimization hyperparameters (AdamW, learning rate, warm-up, and weight decay), and the two-stage training configuration. Architectural details of the Spatial Encoder, Scaler Block, MSFE, and Morph Wrapper are given in Section 3.3. Experiments use frozen Stable Diffusion v1.5 and SDXL backbones, and all baselines are trained on identical datasets for the same number of steps to ensure a fair comparison. FID and CLIP scores are computed on the MS-COCO 2017 validation set. Our code and trained adapter weights will be made publicly available upon acceptance.

REFERENCES

Jiahang Cao, Ziqing Wang, Hanzhong Guo, Hao Cheng, Qiang Zhang, and Renjing Xu. Spiking denoising diffusion probabilistic models. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. doi: 10.1109/wacv57701.2024.00484.

Aryan Das, Koushik Biswas, Swalpa Kumar Roy, Badri Narayana Patro, and Vinay Kumar Verma. Efficient text-guided convolutional adapter for the diffusion model. *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026a. doi: 10.1109/wacv61042.2026.00391.

Aryan Das, Tanishq Rachamalla, Koushik Biswas, Swalpa Kumar Roy, and Vinay Kumar Verma. Uncertainty-aware vision-language segmentation for medical imaging. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026b. doi: 10.1109/wacv61042.2026.00819.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv (Cornell University)*, 2014. doi: arxiv.1410.8516.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv (Cornell University)*, 2020. doi: arxiv.2010.11929.

Demi Guo, Alexander Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*

and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021. doi: 10.18653/v1/2021.acl-long.378.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. doi: 10.18653/v1/2021.emnlp-main.595.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *arXiv (Cornell University)*, 2017. doi: arxiv.1706.08500.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv (Cornell University)*, 2023. doi: arxiv.2302.09778.

Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. In *arXiv (Cornell University)*, 2021. doi: arxiv.2112.05130.

Mohammed Asad Karim and Vinay Kumar Verma. Focus: Forcing in-context object localization through visual support constraints and policy optimization. *arXiv*, 2026. URL https://arxiv.org/abs/2605.31145v1.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback, 2024a. URL https://arxiv.org/abs/2404.07987.

Xirui Li, Charles Herrmann, Kelvin C. K. Chan, Yinxiao Li, Deqing Sun, Chao Ma, and Ming-Hsuan Yang. A simple approach to unifying diffusion-based conditional generation. In *arXiv*, 2024b. URL https://arxiv.org/abs/2410.11439v3.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Lecture Notes in Computer Science*, 2022. doi: 10.1007/978-3-031-20077-9_17.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Lecture Notes in Computer Science*, 2014. doi: 10.1007/978-3-319-10602-1_48.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *arXiv (Cornell University)*, 2017. doi: arxiv.1711.05101.

Siyuan Mei, Mareike Thies, Yan Xia, Yipeng Sun, Fei Wu, Fuxin Fan, Mingxuan Gu, Chengze Ye, Yixing Huang, Vincent Christlein, and Andreas Maier. Vision transformer hook for dense predictions. *Pattern Recognition*, 2026. doi: 10.1016/j.patcog.2026.113818.

Salaheldin Mohamed. Tcig: Two-stage controlled image generation with quality enhancement through diffusion. *arXiv (Cornell University)*, 2024. doi: arxiv.2403.01212.

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*, 2024. doi: 10.1609/aaai.v38i5.28226.

Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/cvpr52729.2023.01769.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. doi: 10.1109/cvpr.2019.00244.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *Advances in Neural Information Processing Systems 36*, 2023a. doi: 10.52202/075280-1862.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *Advances in Neural Information Processing Systems 36*, 2023b. doi: 10.52202/075280-1862.

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. In *arXiv (Cornell University)*, 2021. doi: arxiv.2103.00020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. 2021. doi: arxiv.2102.12092.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv (Cornell University)*, 2022. doi: arxiv.2204.06125.

Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, and Ge Li. Deep image spatial transformation for person image generation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. doi: 10.1109/cvpr42600.2020.00771.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a. doi: 10.1109/cvpr52688.2022.01042.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b. doi: cvpr52688.2022.01042.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494. Curran Associates, Inc., 2022. doi: 10.52202/068431-2643. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

Johannes Schusterbauer, Ming Gui, Frank Fundel, and Björn Ommer. Diff2flow: Training flow matching models via diffusion model alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. doi: 10.1109/cvpr52734.2025.02640.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *arXiv (Cornell University)*, 2020. doi: arxiv.2011.13456.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. doi: 10.1109/iccv51701.2025.01386.

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, Bo Peng, and Yabiao Wang. Unicombine: Unified multi-conditional combination with diffusion transformer. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. doi: 10.1109/iccv51701.2025.01703.

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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/cvpr.2018.00917.

Yuxiang Wei, Ming Liu, Haolin Wang, Ruifeng Zhu, Guosheng Hu, and Wangmeng Zuo. Learning flow-based feature warping for face frontalization with illumination inconsistent supervision. In *Lecture Notes in Computer Science*, 2020. doi: 10.1007/978-3-030-58610-2_33.

Yifeng Xu, Zhenliang He, Shiguang Shan, and Xilin Chen. Ctrlora: An extensible and efficient framework for controllable image generation. In *arXiv (Cornell University)*, 2024. doi: arxiv.2410.09400.

Qingtao Yu, Changlin Song, Minghao Sun, Zhengyang Yu, Vinay Kumar Verma, Soumya Roy, Sumit Negi, Hongdong Li, and Dylan Campbell. Rethinking test time scaling for flow-matching generative models, 2025. URL https://arxiv.org/abs/2511.22242v3.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. doi: iccv51070.2023.00355.

Qinsheng Zhang and Yongxin Chen. Diffusion normalizing flow. In *arXiv*, 2021. URL https://arxiv.org/abs/2110.07579v1.

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 *Advances in Neural Information Processing Systems 36*, 2023. doi: 10.52202/075280-0491.

A Appendix

A.1 Training and Evaluation

Configurations. We evaluate two configurations: UNITY$_{Ind}$, which trains the adapter for each condition from scratch without universal pretraining, and UNITY$_{Pre}$, which follows the two-stage strategy of Section 3.2. Both use a total budget of 100k steps. For UNITY$_{Pre}$, the *Universal Stage* runs for 50k steps on the MS-COCO training set, with all four modalities (Canny, Depth, Sketch, and Segmentation) stacked channel-wise as a 12-channel input in a fixed order. In the *Specialization Stage*, the adapter is initialized from the Universal Stage weights and further finetuned on each modality individually for 50k steps, using MS-COCO and MultiGen-20M combined; for Segmentation, fine-tuning uses MS-COCO only, since MultiGen-20M provides no segmentation maps. To accept single-condition inputs, only the first stage of each spatial encoder is re-initialized for 3 input channels, while all remaining weights are inherited from the Universal Stage.

Optimization. Both SD v1.5 and SDXL backbones, along with their VAE and text encoders, are kept frozen, and only the adapter is trained, at 512×512 resolution for both backbones. We use AdamW (Loshchilov & Hutter, 2017) with a constant learning rate of 1×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, weight decay 1×10^{-2}, and gradient clipping at norm 1.0. Each GPU processes a batch of 2 with 4 gradient-accumulation steps (effective batch size 8). Training uses the Euler discrete noise scheduler with timesteps sampled as $t = (1 - u^3)\,T$, $u \sim \mathcal{U}(0, 1)$, which emphasizes high-noise steps; the objective is the standard noise-prediction loss of Section 3.1. We use fp16 mixed precision in the Universal Stage and bf16 in the Specialization Stage, $M = 3$ sampling points in each Morph Wrapper, and a fixed random seed of 42. All baselines are trained on the same datasets for the same 100k steps.

Evaluation. We report Fréchet Inception Distance (FID) (Heusel et al., 2017) for visual fidelity and CLIP Score (Hessel et al., 2021), computed with CLIP ViT-B/32, for text-image alignment on the MS-COCO 2017 validation set, generating one image for each of its 25,000 image-caption pairs (5,000 images). All images are sampled at 512×512 with the Euler discrete scheduler for 50 steps and a classifier-free guidance scale of 7.5, for both SD v1.5 and SDXL. Further experiments on additional conditioning types are provided in Appendix A.2.

A.2 Additional Experiments

Table 5: Quantitative comparison on additional conditioning modalities. FID↓ and CLIP↑ scores are reported. Best results in **bold**.

Model	Normal Map		Openpose		Color Palette		Dehaze	
	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑
ControlNet (Zhang et al., 2023)	24.46	27.22	25.70	27.08	26.79	26.89	26.12	26.96
Uni-ControlNet (Zhao et al., 2023)	24.29	27.39	25.08	27.33	26.33	27.18	25.86	27.22
CtrLoRA (Xu et al., 2024)	24.94	26.78	26.14	26.55	26.95	26.33	26.69	26.49
UNITY$_{Ind}$	23.85	27.96	24.92	27.39	25.91	27.22	25.01	27.16
UNITY$_{Pre}$	**22.08**	**28.72**	**23.78**	**28.36**	**24.51**	**27.94**	**24.18**	**27.82**

A.2.1 SoTA Results on Stable Diffusion 1.5 with Additional Conditions

We conducted further experiments to validate the generalization capability and robustness of UNITY in various conditioning modalities. We extend our evaluation to four additional conditioning types: **Normal Map**, **Openpose**, **Color Palette**, and **Dehaze**. Following the two-stage training paradigm established in Section 3.2, we train adapters from scratch using the Universal-to-Specialized strategy on these four new conditions and compare against the three top-performing baselines from main experiments: **Uni-ControlNet** (Zhao et al., 2023), **CtrLoRA** (Xu et al., 2024), and **ControlNet** (Zhang et al., 2023). Unlike the main experiments which employed Canny, Depth, Sketch, and Segmentation, this configuration trains exclusively on the new conditioning quartet to assess

UNITY's adaptability to complementary signal types spanning geometric surface properties (Normal Map), human pose articulation (Openpose), content-preserving stylization (Color Palette), and atmospheric degradation modeling (Dehaze).

Conditional Image Generation and Training Configuration: We generate all conditioning inputs using standard annotators for normal maps, Openpose, Color Palette, and Dehaze modalities, applied to both MS-COCO 2017 and MultiGen-20M to form paired training data. MS-COCO provides diverse real-world imagery, while MultiGen-20M supplies large-scale synthetic variation, enabling robust cross-modal learning. All conditioning maps are resized to 512×512 and normalized for compatibility with the Stable Diffusion v1.5 backbone. Training follows a two-stage strategy: a 50k-step Universal Stage that jointly learns across all conditions to build shared representations, followed by a 50k-step Specialization Stage that finetunes each condition independently. This preserves the total 100k-step training budget used in the main experiments while maintaining a constant parameter count (365.25M) and FLOPs (135.82G) irrespective of the number of supported conditioning types.

A.2.2 Results and Discussion

Table 5 presents a detailed quantitative comparison across four additional conditioning modalities. Several clear trends emerge from the FID and CLIP metrics. Among the baselines, Uni-ControlNet consistently performs better than ControlNet and CtrLoRA, although the gains remain modest, with improvements typically in the range of a few tenths of a point for both FID and CLIP, based on the modality.

Both versions of our method provide consistent improvements across all modalities. Even without any pretraining, UNITY$_{Ind}$ already surpasses every baseline in FID, reducing it by up to 0.85, with CLIP comparable or higher. These gains reflect the advantages of our architecture alone. The best results are achieved by the pretrained variant, UNITY$_{Pre}$, which achieves the strongest metric values across all modalities. For Normal Map, it reduces the FID score from 24.29 for Uni-ControlNet to 22.08 and increases the CLIP value from 27.39 to 28.72. For OpenPose, it improves the FID score from 25.08 to 23.78. For Color Palette, it reduces FID from 26.33 to 24.51. For Dehaze, it improves the FID score from 25.86 to 24.18. In all cases, CLIP similarity also increases. The largest gain appears on Normal Map (2.21 FID), suggesting that unified pretraining enhances geometric reasoning and the alignment between the condition and the generated image.

The quantitative results show that unified pretraining followed by a specialization stage provides clear benefits across both geometry-oriented and appearance-oriented modalities. The consistent reductions in FID and increases in CLIP demonstrate enhanced semantic fidelity and reliable controllability across a broad range of conditional image generation tasks.

A.2.3 Additional Qualitative Visualizations

Figures 6, 7, and 8 show samples from UNITY$_{Pre}$ across all eight conditioning types: the four main conditions (Canny Edges, Depth, Sketch, Segmentation) and the four additional conditions of Appendix A.2 (Normal Map, OpenPose, Color Palette, Dehaze). Each row shows the original image, the extracted condition, and three samples from different seeds with the same condition and caption.

Across seeds, outputs follow the spatial structure imposed by the condition, while unspecified attributes such as lighting, color, texture, and background detail vary naturally. Geometry-driven conditions (Canny Edges, Depth, Sketch, Normal Map) closely preserve contours, layout, and surface orientation; e.g., building facades and clock towers keep their architecture while sky and illumination change. Segmentation constrains the placement and extent of scene regions, such as the bed, walls, and furniture in indoor scenes, leaving textures free. OpenPose preserves the person's pose and action (cycling, swinging a tennis racket, kicking a ball), while clothing, surroundings, and viewpoint details differ.

Appearance-oriented conditions behave differently. Color Palette constrains only global color, so layout and geometry vary substantially across seeds while the palette is retained. Dehaze, a restoration-style condition, closely determines the content, yielding near-identical samples that recover clear, high-contrast scenes from hazy inputs. Overall, UNITY$_{Pre}$ adapts to each condition's

level of constraint, giving faithful structural control where the condition is informative and diversity where it is not.

A.3 LIMITATIONS

Although UNITY substantially reduces the cost of multi-condition training, a few aspects remain open. Our experiments focus on UNet-based backbones (SD v1.5 and SDXL), and we have not yet evaluated UNITY with diffusion transformers (DiTs). Adapting it is feasible, since DiTs operate on single-scale token sequences with a fixed embedding size rather than multi-scale feature maps. The MAF outputs could be patchified and linearly projected to the transformer's hidden dimension, similar to the projection layers we already use to match SDXL's feature resolutions, and then injected into the transformer blocks; we leave this extension to future work. The Universal Stage is trained on a predefined set of modalities, so extending UNITY to a new modality currently involves a new universal pretraining phase, as done for the additional conditions in Appendix A.2; supporting incremental modality extension without revisiting this stage is a natural next step. In addition, like prior per-condition methods, deploying every specialized modality involves keeping its finetuned weights, and exploring parameter sharing across specialized adapters could further reduce this footprint. Our evaluation follows standard practice with FID and CLIP Score, and complementing it with condition-specific fidelity metrics and a broader quantitative benchmark for composite conditioning would provide a more complete picture. Finally, as with other adapter-based approaches, UNITY relies on pretrained annotators to produce condition maps and inherits the biases of its frozen diffusion backbone.

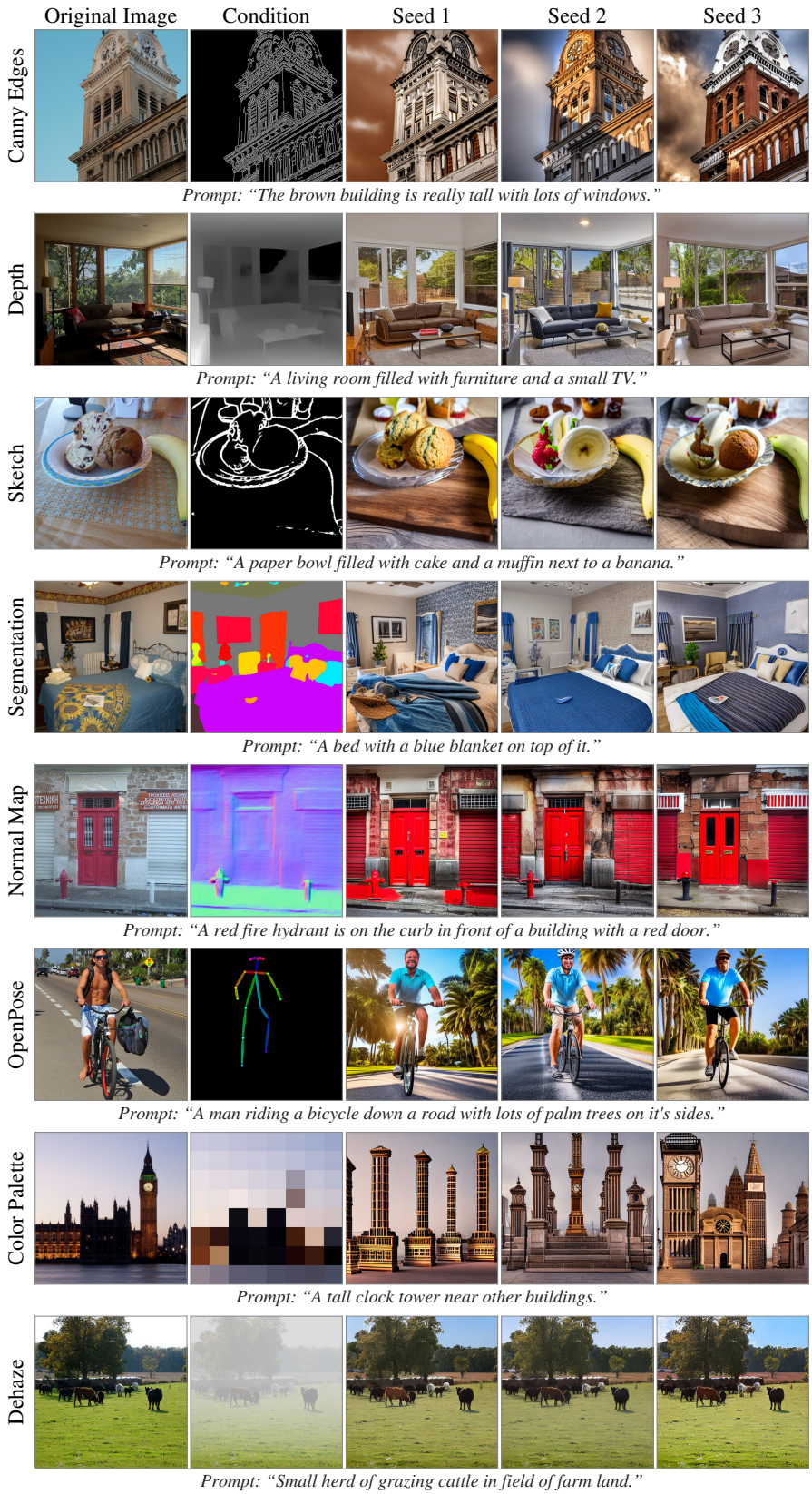


Figure 6: Qualitative results of UNITY$_{Pre}$ across eight conditioning types. Each row shows the original image, the extracted condition, and three samples generated with different random seeds using the caption shown below the row.

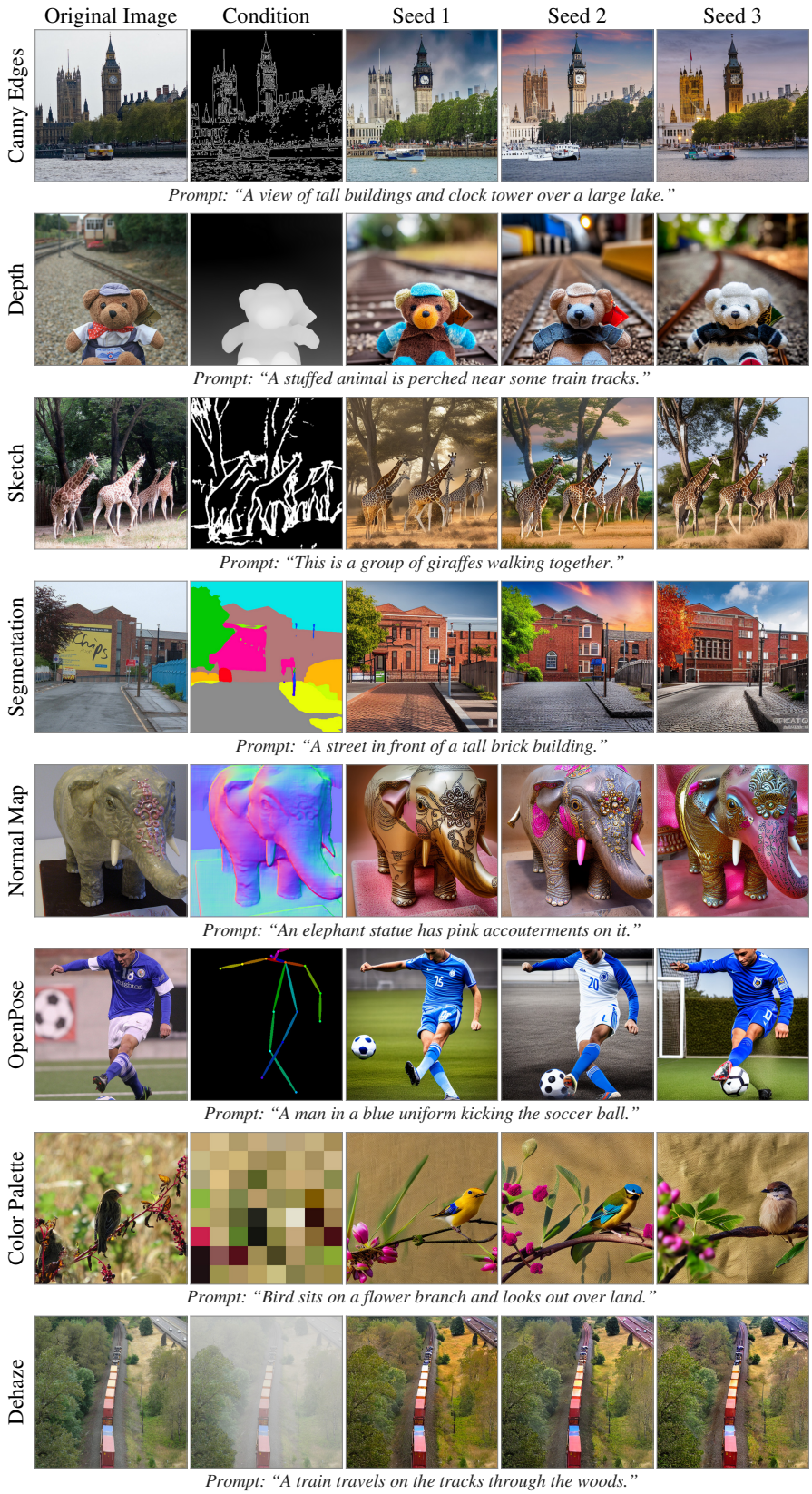


Figure 7: Additional qualitative results of UNITY$_{Pre}$. Samples follow the structure, pose, or layout given by each condition, while unconstrained attributes such as lighting, color, and texture vary across seeds.

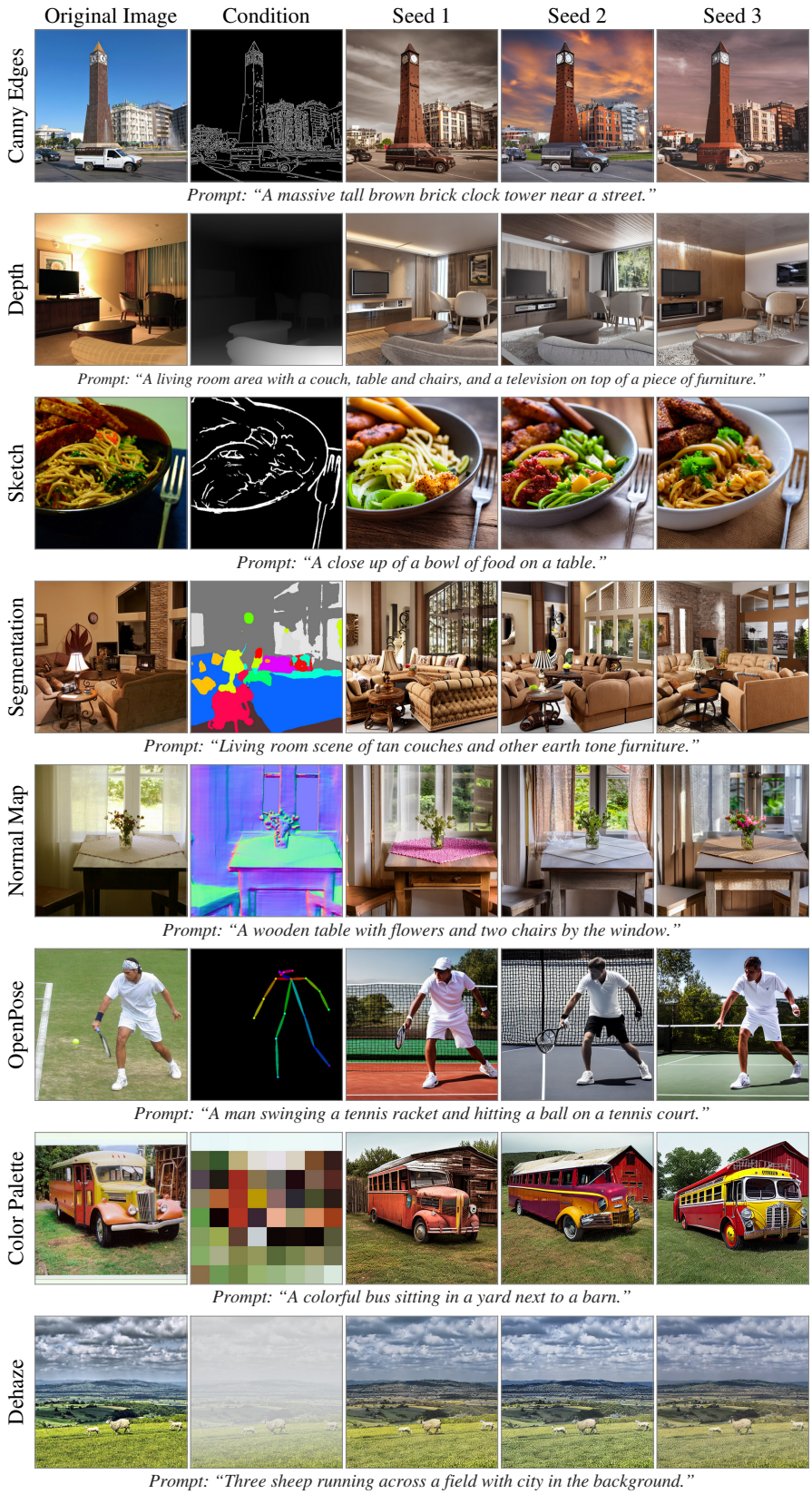


Figure 8: Further qualitative results of UNITY$_{Pre}$. Color Palette constrains only global color, yielding diverse layouts, whereas Dehaze closely determines the output, yielding near-identical samples across seeds.