This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DynFusion: Rethinking Condition Fusion for Adaptive Multi-Conditional Text-to-Image Generation

Zheng Fang[1†], Lichuan Xiang[1†], Xu Cai[2], Bing Wang[3], Bo Yang[3], Hongkai Wen[1*]
[1]*University of Warwick,* [2]*ByteDance,* [3]*Hong Kong Polytechnic University*



Figure 1. **Demonstrations of DynFusion's versatile capabilities. (a)** Controllable image generation under a single visual condition (*e.g.*, subject-driven generation and spatially-aligned tasks). **(b)** Multi-conditional controllable generation. Compared with single-conditional generation, it improves the image quality, controllability and flexibility. Meanwhile, proposed dynamic condition fusion performs better than uniform strategy by eliminating condition redundancies. **(c)** Visualization results under more conditions. **(d)** Quantitative results of our DynFusion. We simultaneously improved both the generation effect and efficiency.

Abstract

Text-to-image diffusion models have achieved remarkable progress, generating visually realistic and semantically coherent images from textual prompts. However, natural language alone lacks the precision required for design-centric applications that demand strict spatial and structural fidelity—particularly when representing complex concepts that integrate multi-level information, such as product or scene design. To address this limitation, controllable diffusion frameworks introduce auxiliary conditions (e.g., depth, edge, or reference images) to guide the generative process. Models like ControlNet[50] and IP-Adapter[48] effectively inject such priors, improving structural or appearance alignment. Yet, real-world design tasks rarely depend on a single type of condition. They often require simultaneous integration of multiple heterogeneous cues—for instance, preserving spatial layout from depth maps, structural outlines from edge maps, and stylistic attributes from reference images. Current approaches either handle only one condition or naively stack multiple ones, resulting in computational inefficiency and conflicting guidance that degrade generation quality.

This multi-condition inconsistency forms a critical bot-

† Equal contribution.
∗ Corresponding author: <hongkai.wen@warwick.ac.uk>.

tleneck for applying diffusion models to real-world design workflows, motivating our proposed framework. We propose a data-driven adaptive condition fusion mechanism for multi-conditional diffusion. Our method introduces a novel condition adaptation module that dynamically selects and fuses subsets of conditions based on the diffusion timestep, task characteristics, and feature injection position. This adaptive strategy harmonizes diverse structural and appearance priors, achieving controllable yet flexible generation in complex design scenarios. Experiments demonstrate significant improvements in fidelity, consistency, and controllability across multi-condition tasks, establishing a new direction for practical, detail-preserving diffusion-based design generation.

1. Introduction

Text-to-image diffusion models have achieved remarkable success in generating photorealistic and semantically coherent images from textual prompts, yet they remain insufficient for precise, instance-level design generation. As our examples illustrate, prompts such as "a girl sleeping in a cozy bedroom" or "a child in a white uniform" describe intent but fail to ensure consistent geometry, identity, or material appearance across conditions—often producing inconsistent poses, confused context, or mismatched references. This limitation arises because natural language alone cannot capture the fine-grained structural and conceptual constraints required in design workflows, where subjects, layouts, and materials must align faithfully with specific references. Consequently, current diffusion models struggle to generalize from abstract text to concrete, reference-grounded design tasks.

To address this limitation, a range of controllable generative frameworks, such as ControlNet [50], InstantID [44], and OmniControl [39], have been developed to integrate diverse reference conditions for guiding diffusion generation. These conditions span from **low-level** structural cues (*e.g.*, depth or edge maps) to **high-level** semantics (*e.g.*, segmentation or identity references), exerting at different diffusion phases. However, most methods rely on manually predefined architectures or fixed timestep to inject conditions into multi-modal blocks (MMBs), without learning *when*, *where*, or *how* each condition should act. As a result, control information is often injected ineffectively and fails to manifest in the generated results, leading to generation **confusion** — results deviate from or inconsistently follow the given conditions, and in some cases, catastrophic **failures** to respect certain conditions or generate broken content. For subject-driven tasks, existing works always **mismatch** on notable details compared to references. FlexControl [8] takes a notable step forward by introducing adaptive condition injection across diffusion steps and network blocks, revealing that effective control depends on both condition

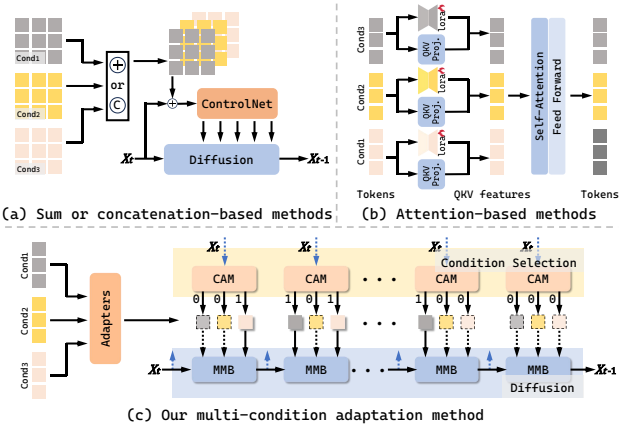


Figure 2. **Comparison of different schemes dealing with multiple conditions.** **(a)** Aggregating various visual conditions into a highly consistent control signal in the temporal domain. **(b)** Noise representation acquires spatial modal knowledge in the latent space through multi-modal attention. **(c)** We employ dynamic condition fusion to precisely match the current demands.

type and contextual relevance within the diffusion process.

Building on the limitations of single-condition diffusion, we observe that real-world design tasks often require multiple complementary conditions to be applied simultaneously. Users may need to preserve space geometry from depth cues, structural outlines from edge maps, and specific object appearances from reference images. Recent extensions such as Cocktail [12], UniControlNet [54], OmniControl2 [40] and UniCombine [42] attempt to support this by stacking additional condition branches or tokens for condition information from various levels. Yet this naive expansion both increases computational cost and fails to account for the heterogeneous nature of **multi-level information**, treating geometric, semantic, and conceptual cues with a uniform fusion strategy that is far from optimal. As a result, the model often suffers from condition interference and control inconsistency, where signals from different levels compete rather than cooperate, leading to structural distortion or semantic drift. Moreover, the redundant conditions further limit scalability and efficiency [7], making such frameworks impractical for real-world multi-conditional applications.

We propose a data-driven adaptive condition fusion mechanism that dynamically determines what, when, and where to control within the diffusion process. As illustrated in Fig. 2, our framework replaces static condition stacking with a multi-condition adaptation module that selectively activates relevant control signals across timesteps and network stages. Unlike prior sum-, concatenation-, or attention-based methods that treat all inputs uniformly, our model learns to assign each condition a context-dependent role—allowing geometric cues, semantic layouts, and appearance references to be injected through distinct, optimally aligned pathways. This not only reduces redundant computation but also enhances controllability and visual

consistency under complex, multi-condition scenarios. Our main contributions are summarized as follows:

- **Novel adaptive multi-condition diffusion framework**. We propose a data-driven condition fusion mechanism that dynamically determines what, when, and where to inject multiple control signals, enabling efficient and coherent guidance across heterogeneous information levels.
- **Comprehensive experimental validation**. Extensive experiments across diverse design and editing tasks demonstrate that our approach achieves superior controllability, efficiency, and visual consistency compared with existing single- and multi-condition diffusion baselines.
- **Analysis of condition activation dynamics**. We systematically investigate how different levels of information - ranging from geometric to semantic and conceptual cues - are dynamically activated and contribute to generation along the diffusion axes of timesteps and network blocks, offering deeper insight into the behavior of adaptive control mechanism.

2. Related Work

Diffusion-based Generative Models. Diffusion models, first proposed by [36], have surpassed Generative Adversarial Network (GAN)-based approaches [13, 43, 46] by offering a more stable and scalable training paradigm [14, 49, 51, 52]. Instead of relying on adversarial optimization—often prone to mode collapse and instability—diffusion models use a likelihood-based denoising process that ensures steady convergence and strong sample diversity. The introduction of Latent Diffusion Models (LDMs) [32] made this approach computationally efficient by operating in a compact latent space without sacrificing fidelity. Building on this foundation, Transformer-based architectures such as DiT [4, 26, 28, 38, 55], Stable Diffusion 3.0/3.5 [6], and FLUX [16] further extend diffusion's capacity to model long-range dependencies and scale to massive multimodal datasets, achieving breakthroughs in compositional reasoning, spatial consistency, and visual realism. Meanwhile, Rectified Flow optimization [20–22] enables faster sampling with fewer denoising steps, improving efficiency without degrading quality. These advances have made diffusion models the foundation of modern creative generation systems.

Controllable Generation. Textual prompts, though expressive, are often too vague to define precise spatial layouts, geometric structures, or stylistic attributes. This limitation drives the development of controllable generation frameworks, which incorporate auxiliary inputs from multiple modalities and abstraction levels—from low-level cues like edges or depth maps to high-level semantic or identity references such as style exemplars or subject images. Representative models such as ControlNet [50], T2I-Adapter [24], and ControlNeXt [27] embed these conditional signals into the generative backbone, enabling spatial alignment and structure preservation. Other approaches like IP-Adapter [48], InstantID [44], and StyleDrop [37] focus on high-level semantic or subject-specific control, integrating reference concepts or styles into the diffusion process to ensure visual and identity consistency. Most existing methods [5, 18, 41] assume a single conditioning source and a unified injection strategy, overlooking the heterogeneity across modalities and abstraction levels. Conditions like edges, depth, layout, or concept carry distinct information and often require tailored architectures or fusion schemes. Recent Transformer-based approaches such as OminiControl [39] and its variants [1, 2, 23, 35] address this by concatenating condition tokens with noise embeddings and applying task-specific LoRA [11], making controllable generation training more consistent with pre-trained text-to-image objectives and thereby improving overall performance.

Multi-Conditional Challenges. To enable precise control, recent studies [17, 47] have extended controllable diffusion models to handle multiple visual conditions. Works such as UniControl [29], UniControlNet [54], and Ctrl-X [19] demonstrate partial success in combining spatially aligned and subject-driven controls. However, most of these methods treat all condition types using a uniform injection mechanism, ignoring the heterogeneity and varying importance of each modality. This one-size-fits-all strategy is far from optimal—low-level cues like edges or depth provide structural guidance, while high-level semantic or stylistic cues require distinct pathways and adaptive weighting.

Extensions such as OminiControl [39] attempt multi-condition fusion by simultaneously activating multiple task-specific LoRA modules within DiT, yet this approach increases computational cost disproportionately to performance gains. Later efforts [40, 53] introduce KV-cache optimizations to mitigate inference latency, but they still overlook conflicts and redundancies among condition signals.

3. Methodology

3.1. Architecture

We exemplify our method on mainstream DiT to ground the discussion. As shown in Fig. 3(a), we reuse the VAE to project condition images into the same latent space as noise tokens, and DiT blocks jointly process both. To minimize the extra computation overhead, we freeze the backbone and train condition-loRAs to adapt DiT blocks for conditional signal extraction, and then inject them into the noise latent space through multi-modal attention (MMA) [40, 42, 53], thus avoiding costly full-parameter tuning.

Our DynFusion introduces a gating mechanism to select optimal condition combinations (Sec. 3.2), enhancing both efficacy and efficiency. We further introduce decoupled multi-modal attention to restrict information flow and preserve conditional signal purity (Sec. 3.3). Fusion-LoRA is proposed to rectify noise distribution under various con-

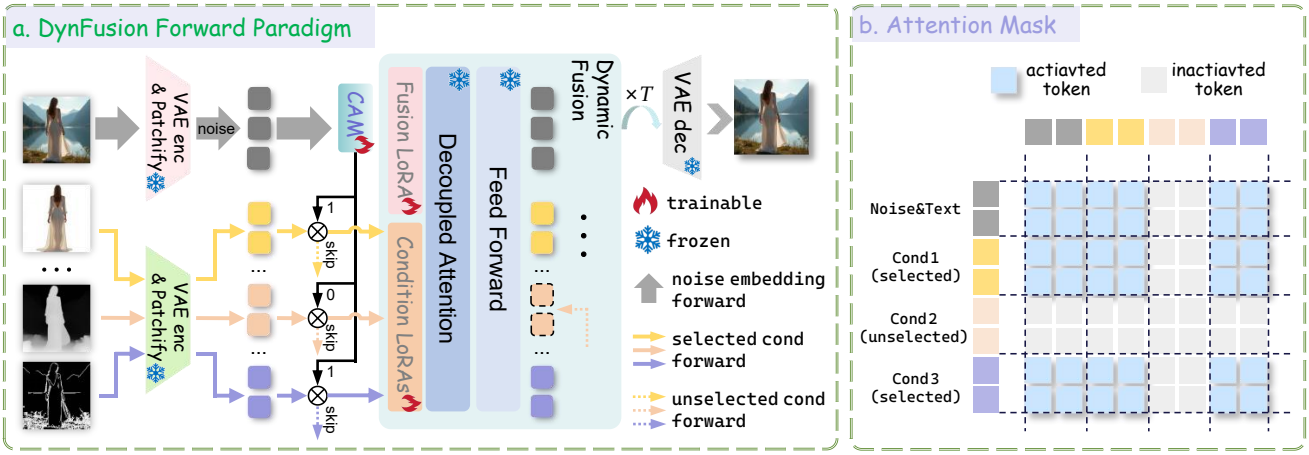


Figure 3. **Overview of our proposed DynFusion framework. (a)** During forward process, the CAM generates instructions based on the noise embeddings to activate the optimal condition combination. **(b)** During training, we incorporate the attention mask with gumbel noise into the softmax calculation of multi-modal attention to filter out the invalid conditions, thus enabling end-to-end training.

dition collections (Sec. 3.4), and a computation-aware loss balances performance with efficiency (Sec. 3.5).

3.2. Adaptive Condition Fusion

Our adaptive condition fusion is realized via the plug-and-play Condition Adaptation Module (CAM). In detail, we equip each multi-modal diffusion transformer (MMDiT) block with CAM for visual condition filtering. As shown in Fig. 3(a), CAM takes noise tokens (**gray squares**) as input ($\mathbf{X} \in \mathbb{R}^{*\times d}$) and predicts the mask $\hat{\mathbf{M}} \in \{0,1\}^n$ over the condition tokens (**yellow, pink and blue squares**).

Specifically, given CAM input tokens are represented as a matrix with dimension of $(* \times d)$, we first apply a decoupled aggregation — using average pooling as the aggregation operator — independently along the token dimension and the embedding dimension. This enables us to extract informative statistics both within each token and across the entire token space. To fuse these two types of features, we employ two separate MLPs to project them into the same latent dimension d_{cam}, which is set to $d/64$ by default. The resulting representations are formulated as:

$$\mathbf{Z}_{\text{global}} = \text{MLP}_{\text{glb}}(\text{Agg}_{\text{seq}}(\mathbf{X}); \Theta_{\text{glb}}), \tag{1}$$

$$\mathbf{Z}_{\text{local}} = \text{MLP}_{\text{loc}}(\text{Agg}_{\text{emb}}(\mathbf{X}); \Theta_{\text{loc}}). \tag{2}$$

The feature aggregated across tokens captures global dependencies and models semantic relationships across regions, whereas the feature aggregated within each token summarizes its overall representation. Both provide complementary cues for token selection [8, 31]. Therefore, we fuse them to obtain a unified global–local feature and feed it into another MLP for mask prediction:

$$\mathbf{Z}^i_{\text{glb-loc}} = \mathbf{Z}^i_{\text{global}} + \mathbf{Z}^i_{\text{local}}, 1 \le i \le d', \tag{3}$$

$$\mathbf{M} = \text{Act}(\text{MLP}_{\text{mask}}(\mathbf{Z}_{\text{glb-loc}}; \Theta_{\text{glb-loc}})), \tag{4}$$

where $\mathbf{M} \in \mathbb{R}^n$, and $\text{Act}(\cdot)$ denotes the activation function.

We consider two choices for $\text{Act}(\cdot)$: (i) Unconstrained selection. When there is no limit on how many conditions can be activated, we use a Sigmoid activation and apply a threshold τ (0.5 by default) to generate the binary mask:

$$\hat{\mathbf{M}}_i = \begin{cases} 1, & \text{if } \mathbf{M}_i > \tau; \\ 0, & \text{otherwise;} \end{cases}, \quad 1 \le i \le n. \tag{5}$$

(ii) Single-condition selection. When selecting only the most essential condition, we use a Softmax activation and apply an argmax operation:

$$\hat{\mathbf{M}}_i = \begin{cases} 1, & \text{if } i = \text{argmax}(\mathbf{M}_i); \\ 0, & \text{otherwise;} \end{cases}, \quad 1 \le i \le n. \tag{6}$$

We compared this two selection mechanism in Tab. 2. Each element in the resulting mask $\hat{\mathbf{M}}$ corresponds to a specific visual condition and acts as a gating signal to screen and filter the conditional controls [8, 31]. As presented in Fig. 3(a), "0" and "1" represent the inactivation and activation of the corresponding control signal from CAM. Inactive condition signals will skip the current MMDiT block and enter the next CAM for the decision-making process (**dotted lines**). Due to the reduced number of condition tokens, the computational complexity of MMDiT is also decreased accordingly. Furthermore, we explore a strategy that involves selecting the most pertinent visual condition. In this way, the computational complexity of MMDiT becomes comparable to single-conditional generation.

The decision-making capability of CAM is learned from a small-scale dataset and evolves dynamically throughout the denoising process. As visualized in Fig. 4, the CAM exhibits distinct activation dynamics across timesteps, conditions, and network blocks, revealing how different control signals contribute throughout the denoising process. The total activation plot **left-corner in** Fig. 4 shows that depth and

MMDiT is mainly composed of MMA and MLP operations. The time complexity of MMA can be formulated as: $\mathcal{O}(n(d^2T+T^2d))$, where n is the set sampling steps, T is the total token account, and d is the embedding dimension. The complexity of MLP is linearly correlated with T.

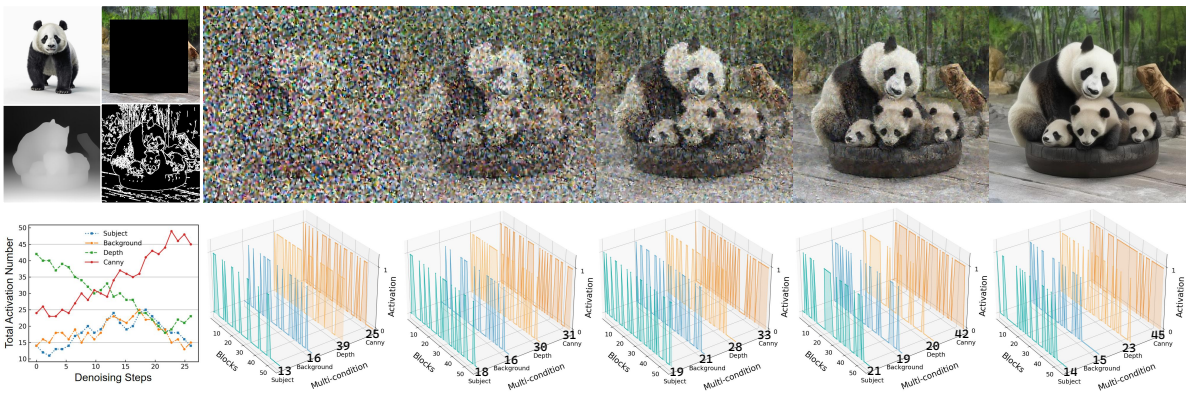


Figure 4. **Visualization of denoising step** at steps 5, 10, 15, 20 and 25, where each pair includes the generated result and condition activation distribution. Subject, background, depth, canny are used to guide image generation.

canny cues maintain consistently higher activation frequencies, implying that structural alignment and edge refinement demand stronger and more persistent control. Their distributions further reveal a complementary temporal behavior—depth activations (**green dash**) dominate the early denoising steps (step 0–10), where the model reconstructs global spatial layout and object placement, while canny activations (**red line**) surge in the later stages (step 15–25), coinciding with contour sharpening and detail recovery in the generated images. In contrast, subject and background activations (**blue and yellow dash**) peak in the middle stages (step 10–20), corresponding to semantic composition where the model integrates foreground entities and contextual backgrounds into a coherent scene.

This evolving activation pattern aligns with the diffusion generation principle: the model first establishes coarse geometry, then integrates semantic information, and finally refines perceptual details. Such stage-aware control demonstrates that CAM adaptively allocates attention to the most relevant conditions at each phase. By synchronizing conditional influence with the natural denoising trajectory, the model avoids over-conditioning and cross-condition conflicts, leading to smoother scene reconstruction, higher spatial consistency, and perceptually sharper results.

Differentiable Learning of CAM. Our target is to perform dynamic condition selection, but it's non-trivial to apply in practice during training. First, the sampling from noise tokens to get the binary decision mask $\hat{\mathbf{M}}$ is non-differentiable, which impedes gradient optimization. To achieve end-to-end training, we incorporate Gumbel Noise [9, 56] into the $\mathrm{Act}(\cdot)$ function. This allows the discrete mask $\hat{\mathbf{M}}$ to be approximately expressed as a differentiable version during the training process.

Parallel Training via Attention Mask. Another obstacle that plagues end-to-end training is that the mask $\hat{\mathbf{M}}$ is unstructured and varies with timesteps and samples. Therefore, simply discarding the condition tokens where $\hat{\mathbf{M}}_i = 0$ would result in non-uniform token numbers for each sample within a batch, making it hard to parallelize the computation. Therefore, we have to keep the token number fixed, while preventing information leakage from the pruned tokens to the remaining ones. However, merely zero-out the unselected condition tokens is not feasible due to the computational characteristics of the Softmax in multi-modal attention. Thus, we remould the attention mask mechanism [31] to eliminate the impact caused by the discarded condition tokens. We first convert the binary mask into the attention mask, as shown in Fig. 3(b), and then apply it to the Softmax. The whole process can be summarized as:

$$\mathbf{P} = QK^\top/\sqrt{d} \in \mathbb{R}^{N\times N}, \tag{7}$$

$$\hat{\mathbf{M}}_{\text{attn}}^{(i,j)} = \begin{cases} 0, & \text{if } \hat{\mathbf{M}}_{\mathcal{C}(i)} \wedge \hat{\mathbf{M}}_{\mathcal{C}(j)} = 0; \\ 1, & \text{otherwise;} \end{cases}, 1 \le i,j \le N, \tag{8}$$

$$\tilde{\mathbf{A}}^{(i,j)} = \frac{\exp\left(\mathbf{P}^{(i,j)}\right)\hat{\mathbf{M}}_{\text{attn}}^{(i,j)}}{\sum_{k=1}^{N}\exp\left(\mathbf{P}^{(i,j)}\right)\hat{\mathbf{M}}_{\text{attn}}^{(i,j)} + \varepsilon}, 1 \le i,j \le N. \tag{9}$$

Following Eqs. (7) to (9), we reconstruct the attention map. Here, $\mathcal{C}(\cdot)$ is used to convert the token index to the condition index, and ε is a hyperparameter used to prevent underflow. The masked attention map $\tilde{\mathbf{A}}$ calculated by Eq. (9) is equivalent to the attention map obtained by removing the unselected tokens. Meanwhile, it maintains the shape of matrices, enabling to perform parallel training.

Remark: The CAM is lightweight, comprising less than 1% of total parameters. During inference, inactive conditions are skipped, reducing computation while adaptively selecting effective condition combinations to enhance generation quality. The detailed procedure is provided in Algorithm 1 (in Supplementary).

3.3. Decoupled Multi-Modal Attention.

Our framework isolates the conditional branch and the denoising branch into computation-independent modules, and the denoising branch's essence is to facilitate the noise&text tokens to learn spatial semantic information from the condition tokens during the denoising process, rather than conducting information exchange between the condition tokens. For example, when noise embedding X serves as

query, the attention map of decoupled multi-modal attention (DMMA) is computed between X and all conditions:

$$\begin{aligned}&\mathrm{DMMA}(Q=X^q, K/V=[C_T^{k/v}, X^{k/v}, C_{V_{1:n}}^{k/v}])\\ =&\mathrm{Softmax}(\frac{1}{\sqrt{d}}X^q[C_T^k, X^k, C_{V_{1:n}}^k]^\top)[C_T^v, X^k, C_{V_{1:n}}^v].\end{aligned} \tag{10}$$

Furthermore, compared to noise tokens, conditional tokens remain agnostic to the diffusion process [40, 53]:

$$\begin{aligned}&\mathrm{DMMA}(Q=C_{V_i}^q, K=C_{V_i}^k, V=C_{V_i}^v)\\ =&\mathrm{Softmax}(\frac{1}{\sqrt{d}}C_{V_i}^q[C_{V_i}^k]^\top)C_{V_i}^v.\end{aligned} \tag{11}$$

This paradigm enables the noise&text tokens to fuse the conditional features effectively and ensure that there will be no confusion of information among the conditions.

3.4. Fusion-LoRA Module

Since condition-LoRA branches are trained independently, when multiple visual conditions are activated simultaneously, relying solely on the Softmax operation in Eq. (10) to balance the attention score distributions of multiple condition embeddings may lead to suboptimal feature fusion results [42]. Furthermore, within our framework, the combinations of conditions and the quantities involved dynamically vary throughout the generation process, which further exacerbates this situation. To address the above issues, we introduced a Fusion-LoRA module in the denoising branch to rectify noise latent feature distribution, thus better integrating multiple control information.

3.5. Loss Functions

We use flow-matching [6] loss as the training objective for DynFusion. Given the training dataset $\mathcal{D}$, which consists of tuples of the original image x, a set of visual condition images C_V, and text conditions C_T. The mathematical expression of the loss function is as follows:

$$\mathcal{L}_{\mathrm{diff}}(\Theta)=\mathbb{E}_{t,\epsilon\sim\mathcal{N}(0,I)}\left\|v_\Theta(z,t,C_T,C_V)-u_t(z|\epsilon)\right\|_2^2, \tag{12}$$

here, z denotes the position of the data points at timestep t, ϵ is the predicted noise vector, v_Θ refers to the velocity learned by the model, and u_t represents the actual vector field of the reference path.

Additionally, to facilitate the adjustment of the condition sparsity, we also introduce the sparsity loss [8, 55] to regulate condition overhead to the desired sparsity λ:

$$\mathcal{L}_{\mathrm{sps}}=\frac{1}{|\mathcal{D}_{\mathrm{bs}}|}\sum_{d\in\mathcal{D}_{\mathrm{bs}}}(\frac{\mathrm{F}_{\mathrm{dynamic}}^{t_d}}{\mathrm{F}_{\mathrm{uniform}}}-\lambda)^2, \tag{13}$$

where $\mathcal{D}_{\mathrm{bs}}$ means a batch of sample tuples, $t_d\in[0,T]$ is the uniformly sampled timestep for sample d, $\mathrm{F}_{\mathrm{dynamic}}^{t_d}$ and $\mathrm{F}_{\mathrm{uniform}}$ are the floating-point operations (FLOPs) consumed when processing the conditional image under adaptive condition fusion and uniform fusion setting, respectively. We combine $\mathcal{L}_{\mathrm{diff}}$ and $\mathcal{L}_{\mathrm{sps}}$ as optimization object:

$$\mathcal{L}_\theta=\mathcal{L}_{\mathrm{diff}}+\alpha\cdot\mathcal{L}_{\mathrm{sps}}, \tag{14}$$

where α serves as the weight of the sparsity loss and is set to 1.0 by default.

4. Experiment

We evaluate DynFusion against state-of-the-art methods across different tasks: Multi-Spatial generation, Subject-Insertion generation, Subject-Spatial generation, and *etc*. Furthermore, we conduct extensive ablation experiments to prove the efficacy and efficiency of our method. Implementation details are listed in Sec. A3 (in Supplementary).

4.1. Main Results

Multi-Spatial Generation. This experiment addresses the limited spatial representation capability of using a single condition in spatially aligned generative tasks. To improve spatial consistency and visual fidelity, multiple spatial conditions are introduced in a complementary manner, providing stronger geometric and structural constraints during image generation. As shown in Tab. 1 and Fig. 5, Our method surpasses existing approaches in image quality and controllability, preserving structures more accurately. The generated buildings align well with the depth map, maintaining correct hollow geometry. Additionally, fine details such as fences follow the canny edges precisely, avoiding the wall–fence confusion seen in prior methods.

Subject-Insertion Generation. Compared with subject-driven generation that only uses the reference target as conditional control, subject-insertion generation inserts the reference subject into the masked region of the background image, thus better meeting the actual production requirements. As shown in Tab. 1, our DynFusion demonstrates superior performance compared to other advanced methods in image quality and generative consistency. Meanwhile, it can be seen from Fig. 6, our method can better preserve the shape and integrity of the reference subject, especially when the subject and background content are abundant and the structure is complex. Moreover, our method does not simply paste the subject image onto the masked area in a rough manner, but integrates it into the background in a highly consistent and harmonious way. Additionally, our method does not require test-time tuning as traditional customization methods [15, 33] do.

Subject-Spatial Generation. Try-on and place-on tasks present significant challenges in controllable generation, as they require the reference subject to be accurately generated and seamlessly integrated into specific scene. To evaluate performance, we conduct experiments using two common spatial conditions: depth maps and canny edges. As shown in the Subject-Depth and Subject-Canny experiments in Tab. 1, DynFusion consistently delivers superior generative performance while achieving an effective balance between quality and computational efficiency. As shown in Figs. 7 and 8, DynFusion demonstrates superior adaptability, accurately adjusting subject morphology according to spatial conditions and text prompts. It also maintains strong subject consistency, precisely locating subjects without mixing

Task	Method	Base Model	Generative Quality FID ↓	SSIM ↑	Controllability F1 ↑	MSE ↓	Subject Consistency CLIP-I ↑	DINO ↑	Text Consistency CLIP-T ↑	Inference Overhead Params ↓	FLOPs ↓	Speed ↑
Multi-Spatial	UniControl	SD1.5	44.17	0.32	0.07	1346.02	-	-	30.28	-	-	**5.46it/s**
	UniControlNet		20.96	0.28	0.09	1231.06	-	-	32.74	-	-	4.36it/s
	ControlNet-Union		54.20	0.25	0.07	1853.42	-	-	30.72	3302M	10.65T	**3.15it/s**
	OminiControl	FLUX.1	10.94	0.54	0.18	524.86	-	-	32.49	**29M**	16.92T	1.46it/s
	UniCombine		7.29	0.61	**0.24**	187.66	-	-	**33.48**	44M	16.21T	1.42it/s
	Ours		**6.52**	**0.66**	**0.24**	**158.97**	-	-	33.44	47M	**8.27T**	2.02it/s
Subject-Insertion	ObjectStitch	SD1.5	26.86	0.37	-	-	93.05	82.34	32.25	-	-	1.46it/s
	AnyDoor	SD2.1	26.07	0.37	-	-	94.88	86.04	32.55	-	-	**6.83it/s**
	OminiControl		11.27	0.66	-	-	95.38	87.64	33.08	**29M**	16.92T	1.46it/s
	UniCombine	FLUX.1	5.14	0.76	-	-	96.95	92.54	33.12	44M	16.21T	1.42it/s
	Ours		**4.53**	**0.80**	-	-	**97.21**	**93.14**	**33.21**	47M	**7.76T**	**2.09it/s**
Subject-Depth	Ctrl-X	SDXL	52.37	0.36	-	2644.90	78.08	50.83	30.20	-	-	**2.64it/s**
	ControlNet w. IP-Adapter		29.93	0.34	-	1295.80	80.41	62.26	32.94	4214M	**5.41T**	**3.26it/s**
	OminiControl	FLUX.1	10.74	0.51	-	501.76	88.96	84.73	32.80	**29M**	16.92T	1.46it/s
	UniCombine		6.92	0.52	-	206.58	93.79	90.41	33.26	44M	16.21T	1.42it/s
	Ours		**6.21**	**0.56**	-	**194.25**	**94.52**	**90.70**	**33.48**	47M	7.96T	2.04it/s
Subject-Canny	Ctrl-X	SDXL	47.89	0.36	0.05	-	79.35	54.31	30.34	-	-	**2.64it/s**
	ControlNet w. IP-Adapter		30.38	0.38	0.09	-	79.80	60.19	32.85	4214M	**5.41T**	**3.26it/s**
	OminiControl	FLUX.1	11.02	0.50	0.17	-	91.52	86.75	32.61	**29M**	16.92T	1.46it/s
	UniCombine		6.41	0.57	0.22	-	94.76	92.24	33.27	44M	16.21T	1.42it/s
	Ours		**5.72**	**0.64**	**0.25**	-	**95.33**	**92.87**	**33.32**	47M	8.20T	2.02it/s

Table 1. **Quantitative comparison of our method with existing approaches** on Multi-Spatial, Subject-Insertion, Subject-Depth, and Subject-Canny conditional generative tasks. The **bold** and underlined figures represent the optimal and sub-optimal results, respectively. In the last column, "Params↓" and "FLOPs↓" are additional introduced parameters and floating-point operations for processing visual conditions on FLUX, and "Speed↑" denotes diffusion iterations per second, which is the average on test set.

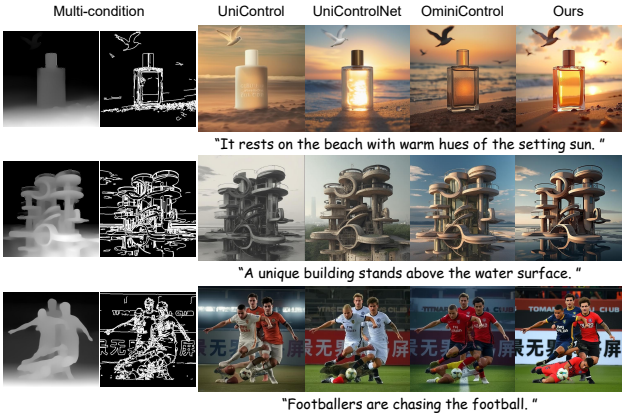


Figure 5. Qualitative comparison on Multi-Spatial generation.



Figure 6. Qualitative comparison on Subject-Insertion generation.

appearance features with spatial semantics. Moreover, it preserves edge fidelity and fine details, achieving coherent alignment between spatial structure and textual intent.

4.2. Ablation Studies

In this section, we conduct ablation study on the Subject-Insertion generative task, while more experiments on other tasks are provided in Sec. A4 (in Supplementary).

Effect of Adaptive Condition Fusion. To assess the impact of adaptive condition fusion, we compare our dynamic fusion strategy with two baselines: (i) uniformly activating all visual conditions and (ii) activating only a single essential condition, while keeping all other settings fixed. As shown in Tab. 2 and Fig. A1 (in Supplementary), DynFusion achieves higher generation quality with lower computational cost. This confirms that redundant or conflicting condition signals can degrade generation by confusing subject features and hindering accurate appearance restoration—issues effectively mitigated by our adaptive fusion mechanism. Moreover, the decision-making behavior of the Condition Adaptation Module (CAM) evolves dynamically throughout the denoising process. As visualized in Fig. 4, CAM exhibits distinct activation patterns across timesteps, conditions, and MMDiT blocks, revealing how each control signal contributes at different denoising stages. Depth and canny cues show persistently high activation frequencies, reflecting their critical role in spatial alignment and edge refinement. Specifically, depth activations dominate early steps (0–10) to reconstruct global structure, canny activations surge in later steps (15–25) for contour and detail recovery, while subject and background cues peak mid-process (10–20) to integrate semantic content coherently. This dynamic adaptation enables flexible, context-aware condition fusion—balancing efficiency with precise, structure-consistent generation.

Effect of Decoupled Multi-Modal Attention. To verify the effectiveness of decoupled attention, we replace DMMA with MMA [39] and CMMA [42] for training and perfor-

UniCombine is the training-based version for better generation performance throughout all the experimental stages.



Figure 7. Qualitative comparison on Subject-Depth generation.

Method	FID ↓	SSIM ↑	CLIP-I ↑	DINO ↑	CLIP-T ↑	FLOPs ↓
Ours.Uniform	5.06	0.76	96.92	92.71	33.15	15.10T
Ours.Sole	4.89	0.78	97.08	92.96	**33.24**	**7.56T**
Ours.Free	**4.53**	**0.80**	**97.21**	**93.14**	33.21	7.76T

Table 2. **Quantitative ablation of adaptive condition fusion strategy** on Subject-Insertion task. "Uniform" means activating all conditions uniformly. "Sole" and "Free" indicate selecting the conditions with softmax and sigmoid function, respectively.

Method	FID ↓	SSIM ↑	CLIP-I ↑	DINO ↑	CLIP-T ↑	AttnOps ↓
Ours w. MMA	5.19	0.78	96.79	92.40	33.08	2.74T / 4.50T
Ours w. CMMA	4.75	**0.81**	97.13	92.99	33.11	2.41T / 3.76T
Ours w. DMMA	**4.53**	0.80	**97.21**	**93.14**	**33.21**	**1.77T / 2.66T**

Table 3. **Quantitative ablation of decoupled multi-modal attention** on Subject-Insertion task. "AttnOps↓" is short for attention operations. Here, the former part is the actual overhead, while the latter is the theoretical overhead under uniform setting.

mance evaluation. As shown in Tab. 3 and Fig. A2 (in Supplementary), our framework achieves superior performance and fewer attention operations. This indicates that if the condition absorbs information from other conditions or noise, its semantic content will become diluted and entangled. Consequently, during adaptive fusion, features that should be shielded will be contained in the activated conditions, thereby interfering with the intended control effect.

Effect of Fusion-LoRA. We evaluate the necessity of Fusion-LoRA, with results presented in Tab. 4 and Fig. A3 (in Supplementary). Removing the Fusion-LoRA module leads to a substantial performance drop. This demonstrates that Fusion-LoRA plays a significant role in modulating the conditional feature aggregation operation. Especially under the effect of our adaptive condition fusion strategy, the categories and quantities of activated conditions are dynamically adjusted over timesteps and samples. Therefore, Fusion-LoRA is crucial for enabling noise embeddings to adapt to these dynamic adjustment.

Different Condition Sparsity. We also explore the performance of DynFusion under different levels of condition sparsity. We control the condition sparsity by adding sparsity loss and tuning the value of λ. As presented in Tab. 5 and Fig. A4 (in Supplementary), when the sparsity is set

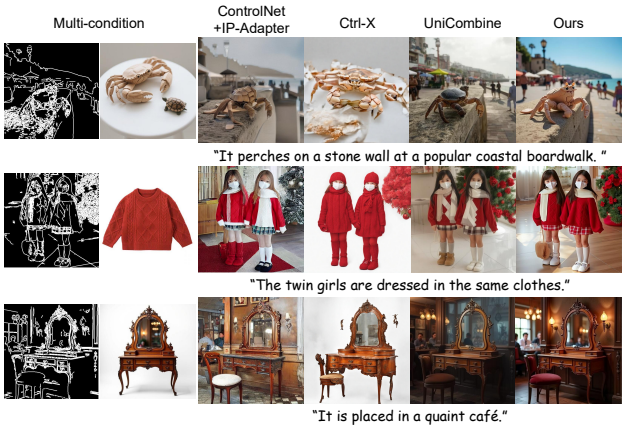


Figure 8. Qualitative comparison on Subject-Canny generation.

Method	FID ↓	SSIM ↑	CLIP-I ↑	DINO ↑	CLIP-T ↑
Ours w/o. Fusion-LoRA	5.93	0.73	96.19	89.42	33.10
Ours w. Fusion-LoRA	**4.53**	**0.80**	**97.21**	**93.14**	**33.21**

Table 4. **Quantitative ablation of Fusion-LoRA component** on Subject-Insertion task. Fusion-LoRA is capable of modulating different quantities and combinations of conditional signals, thereby providing better assistance in generating.

Method	FID ↓	SSIM ↑	CLIP-I ↑	DINO ↑	CLIP-T ↑	Speed ↑
Ours.Uniform	5.06	0.76	96.92	92.71	33.15	1.48it/s
Ours (×0.7)	4.85	**0.78**	97.05	92.99	33.14	1.84it/s
Ours (×0.5)	**4.77**	**0.78**	**97.14**	**93.08**	**33.21**	2.11it/s
Ours (×0.3)	5.19	0.75	96.79	92.52	32.98	**2.49it/s**

Table 5. **Quantitative ablation of condition sparsity** on Subject-Insertion task.

to 50%, the generative quality and controllability of the model are the best. Reducing sparsity to 30% weakens control strength due to insufficient conditioning signals, leading to a notable performance drop. When increasing sparsity to 70%, the performance slightly declines compared to 50%, indicating that excessive control signals introduce redundancy. When all conditions are further activated, performance degrades substantially, suggesting that excessive conditional controls cause conflicts and further redundancy, which have a negative impact on the model. Therefore, dynamically injecting the conditions is a promising direction for optimizing controllable diffusion generation.

5. Conclusion

We propose DynFusion, an adaptive condition fusion framework that replaces static fusion with a lightweight, plug-and-play gating mechanism. By dynamically selecting condition combinations for each MMDiT block and sample, DynFusion mitigates condition conflicts, removes redundant computations, and enhances generative reliability. The Condition Adaptation Module autonomously manages this process, achieving efficient and coherent fusion across heterogeneous conditions. Our results demonstrate both improved performance and scalability, offering a practical direction for future multi-conditional control research.

References

[1] Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18434–18443, 2025. 3

[2] Ke Cao, Jing Wang, Ao Ma, Jiasong Feng, Zhanjie Zhang, Xuanhua He, Shanyuan Liu, Bo Cheng, Dawei Leng, Yuhui Yin, et al. Relactrl: Relevance-guided efficient control for diffusion transformers. *arXiv preprint arXiv:2502.14377*, 2025. 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 13

[4] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12501–12511, 2025. 3

[5] Zhaoli Deng, Kaibin Zhou, Fanyi Wang, and Zhenpeng Mi. Repcontrolnet: Controlnet reparameterization. *arXiv preprint arXiv:2408.09240*, 2024. 3

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis, march 2024. *URL http://arxiv. org/abs/2403.03206*, 2024. 3, 6

[7] Zheng Fang and Bo Yin. A low functional redundancy-based network slimming method for accelerating deep neural networks. *Alexandria Engineering Journal*, 119:437–450, 2025. 2

[8] Zheng Fang, Lichuan Xiang, Xu Cai, Kaicheng Zhou, and Hongkai Wen. Flexcontrol: Computation-aware conditional control with differentiable router for text-to-image generation. In *Forty-second International Conference on Machine Learning*. 2, 4, 6

[9] Charles Herrmann, Richard Strong Bowen, and Ramin Zabih. Channel selection using gumbel softmax. In *European conference on computer vision*, pages 241–257. Springer, 2020. 5

[10] 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. *Advances in neural information processing systems*, 30, 2017. 13

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3

[12] Minghui Hu, Jianbin Zheng, Daqing Liu, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-Jen Cham. Cocktail: Mixing multi-modality control for text-conditional image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[13] Kaitao Huang, Yan Yan, Jing-Hao Xue, and Hanzi Wang. Warpgan: Warping-guided 3d gan inversion with style-based novel view inpainting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 3

[14] Rui Huang, Shitong Shao, Zikai Zhou, Pukun Zhao, Hangyu Guo, Tian Ye, Lichen Bai, Shuo Yang, and Zeke Xie. Diffusion dataset condensation: Training your diffusion model faster with less data. *arXiv preprint arXiv:2507.05914*, 2025. 3

[15] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 6

[16] Black Forest Labs. Flux: Official inference repository for flux.1 models, 2024. Accessed: 2024-11-12. 3, 13

[17] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024. 3

[18] Jiang Lin, Xinyu Chen, Song Wu, Zhiqiu Zhang, Jizhi Zhang, Ye Wang, Qiang Tang, Qian Wang, Jian Yang, and Zili Yi. Freecontrol: Efficient, training-free structural control via one-step attention extraction. *arXiv preprint arXiv:2511.05219*, 2025. 3

[19] 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. *Advances in Neural Information Processing Systems*, 37: 128911–128939, 2024. 3

[20] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. 3

[21] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[22] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[23] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025. 3

[24] 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. 3

[25] Yanjie Pan, Qingdong He, Zhengkai Jiang, Pengcheng Xu, Chaoyi Wang, Jinlong Peng, Haoxuan Wang, Yun Cao, Zhenye Gan, Mingmin Chi, et al. Pixelponder: Dynamic patch adaptation for enhanced multi-conditional text-to-image generation. *arXiv preprint arXiv:2503.06684*, 2025. 12

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF Inter-*

national Conference on Computer Vision, pages 4195–4205, 2023. 3

[27] 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. 3

[28] Yifan Pu, Zhuofan Xia, Jiayi Guo, Dongchen Han, Qixiu Li, Duo Li, Yuhui Yuan, Ji Li, Yizeng Han, Shiji Song, et al. Efficient diffusion transformer with step-wise dynamic attention mediators. In *European Conference on Computer Vision*, pages 424–441. Springer, 2024. 3

[29] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 3

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 13

[31] Yongming Rao, Zuyan Liu, Wenliang Zhao, Jie Zhou, and Jiwen Lu. Dynamic spatial sparsification for efficient vision transformers and convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10883–10897, 2023. 4, 5

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 12

[33] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 6

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 12

[35] D She, Mushui Liu, Jingxuan Pang, Jin Wang, Zhen Yang, Wanggui He, Guanghao Zhang, Yi Wang, Qihan Huang, Haobin Tang, et al. Customvideox: 3d reference attention driven dynamic adaptation for zero-shot customized video diffusion transformers. *arXiv preprint arXiv:2502.06527*, 2025. 3

[36] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[37] Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, et al. Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems*, 36:66860–66889, 2023. 3

[38] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3, 12

[39] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024. 2, 3, 7, 13

[40] Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280*, 2025. 2, 3, 6, 13

[41] Long Tang, Tingting Chai, Zheng Zhang, Miao Zhang, and Xiangqian Wu. Palmdiff: When palmprint generation meets controllable diffusion model. *IEEE Transactions on Image Processing*, 2025. 3

[42] Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. Unicombine: Unified multi-conditional combination with diffusion transformer. *arXiv preprint arXiv:2503.09277*, 2025. 2, 3, 6, 7, 13

[43] Jianyuan Wang, Lalit Bhagat, Ceyuan Yang, Yinghao Xu, Yujun Shen, Hongdong Li, and Bolei Zhou. Spatial steerability of gans via self-supervision from discriminator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9493–9507, 2024. 3

[44] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 3

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 13

[46] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024. 3

[47] Hongji Yang, Wencheng Han, Yucheng Zhou, and Jianbing Shen. Dc-controlnet: Decoupling inter-and intra-element conditions in image generation with diffusion models. *arXiv preprint arXiv:2502.14779*, 2025. 3

[48] 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, 3

[49] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025. 3

[50] 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, 3

[51] Rongchao Zhang, Yu Huang, Yiwei Lou, Weiping Ding, Yongzhi Cao, and Hanpin Wang. Synergistic attention-guided cascaded graph diffusion model for complementarity determining region synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):11875–11886, 2024. 3

[52] Rongchao Zhang, Yu Huang, Yiwei Lou, Yi Xin, Haixu Chen, Yongzhi Cao, and Hanpin Wang. Exploit your latents: Coarse-grained protein backmapping with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1111–1119, 2025. 3

[53] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19513–19524, 2025. 3, 6

[54] 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. 2, 3

[55] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic diffusion transformer. *arXiv preprint arXiv:2410.03456*, 2024. 3, 6

[56] Wangbo Zhao, Jiasheng Tang, Yizeng Han, Yibing Song, Kai Wang, Gao Huang, Fan Wang, and Yang You. Dynamic tuning towards parameter and inference efficiency for vit adaptation. *Advances in Neural Information Processing Systems*, 37:114765–114796, 2024. 5