DivControl: Knowledge Diversion for Controllable Image Generation

Yucheng Xie[1,2], **Fu Feng**[1,2], **Ruixiao Shi**[1,2], **Jing Wang**[1,2*], **Yong Rui**[1,2], **Xin Geng**[1,2*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
{xieyc, fufeng, eric_xiao, wangjing91, xgeng}@seu.edu.cn

Abstract

Diffusion models have advanced from text-to-image (T2I) to image-to-image (I2I) generation by incorporating structured inputs such as depth maps, enabling fine-grained spatial control. However, existing methods either train separate models for each condition or rely on unified architectures with entangled representations, resulting in poor generalization and high adaptation costs for novel conditions. To this end, we propose **DivControl**, a decomposable pretraining framework for unified controllable generation and efficient adaptation. DivControl factorizes ControlNet via SVD into basic components—pairs of singular vectors—which are disentangled into condition-agnostic learngenes and condition-specific tailors through knowledge diversion during multi-condition training. Knowledge diversion is implemented via a dynamic gate that performs soft routing over tailors based on the semantics of condition instructions, enabling zero-shot generalization and parameter-efficient adaptation to novel conditions. To further improve condition fidelity and training efficiency, we introduce a representation alignment loss that aligns condition embeddings with early diffusion features. Extensive experiments demonstrate that DivControl achieves state-of-the-art controllability with 36.4× less training cost, while simultaneously improving average performance on basic conditions. It also delivers strong zero-shot and few-shot performance on unseen conditions, demonstrating superior scalability, modularity, and transferability.

1 Introduction

Diffusion models have demonstrated remarkable performance in text-to-image (T2I) generation, with models such as DALL-E 3 (Ramesh et al. 2022), Stable Diffusion 3 (Esser et al. 2024), and Midjourney (Midjourney 2022) generating images comparable to human-created artwork from natural language prompts (Balaji et al. 2022; Feng et al. 2025d; Zhang, Rao, and Agrawala 2023). Their ability to produce high-fidelity, semantically aligned outputs has positioned diffusion models as a foundation for controllable generation. To support finer-grained and more deterministic control, recent efforts have extended conditions beyond text to structured visual inputs—such as depth and segmentation maps—enabling image-to-image (I2I) generation with explicit spatial guidance (Zhang, Rao, and Agrawala 2023; Zavadski, Feiden, and Rother 2024; Mou et al. 2024).

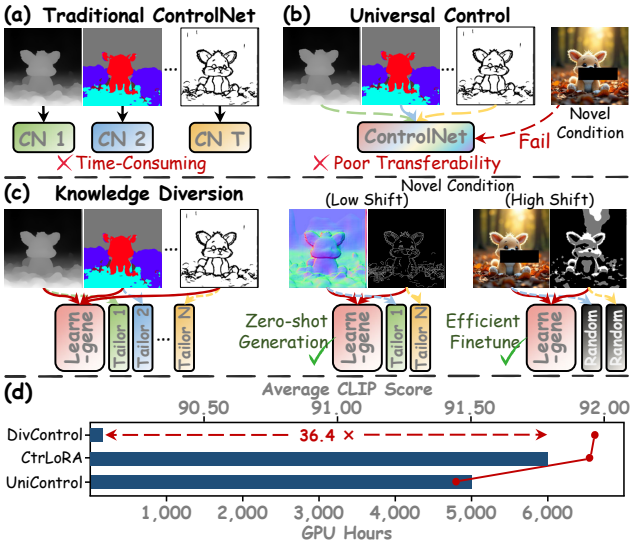

Figure 1: (a) Traditional ControlNet requires training a dedicated model for each control type, leading to substantial computational overhead. (b) Universal control approaches aim to unify all conditions within a single model, but exhibit poor generalization to unseen tasks. (c) DivControl addresses this by introducing knowledge diversion to disentangle condition-agnostic and condition-specific representations during training, enabling unified control and zero-shot transfer. (d) This modular design reduces training cost by over **36.4**× (165 vs. 6000 GPU hours) while achieving superior controllable generation quality.

However, training separate diffusion models for each control condition is computationally prohibitive. For instance, ControlNet (Zhang, Rao, and Agrawala 2023) requires over 600 A100 GPU hours on 3 million images to support a single CANNY condition, with other modalities demanding even more resources (Xu et al. 2024). To improve efficiency and scalability, recent works propose unified frameworks for multi-condition control (Qin et al. 2023; Tan et al. 2024; Wang et al. 2025), aiming to handle diverse conditions within a single model (Figure 1a, b). However, these ap-

*Corresponding authors

proaches struggle to generalize to unseen or heterogeneous conditions, as jointly learned representations tend to be entangled, hindering adaptation to novel control conditions.

While CtrLoRA (Xu et al. 2024) improves transferability by assigning a dedicated LoRA (Hu et al. 2022) to each condition during training, thereby shaping a more adaptable ControlNet, it remains computationally intensive due to large-scale multi-condition pretraining (Feng and Zhang 2023). More critically, its rigid separation mechanism fails to account for intrinsic inter-condition correlations, leading to suboptimal performance, limiting modular reuse and hindering generalization to unseen conditions.

Recently, knowledge diversion (Xie et al. 2025) was introduced to explicitly disentangle task-agnostic and task-specific knowledge by applying Singular Value Decomposition (SVD) to factorize network weight matrices into shared learngenes and task-specific tailors, thereby improving modular reuse and cross-task transferability. Building on this principle, we propose DivControl, which brings knowledge diversion into controllable image generation. By factorizing ControlNet into shared learngenes and lightweight tailors during training, DivControl supports unified generation across diverse conditions and enables efficient adaptation to novel conditions with minimal overhead.

DivControl applies SVD to decompose each weight matrix into basic components—pairs of singular vectors—which are modularly assigns them to shared learngenes or condition-specific tailors for structured control. To improve parameter sharing and generalization, DivControl replaces the binary gate (Xie et al. 2025; Xu et al. 2024) with a dynamic gate. Specifically, each condition is described by a textual instruction, which is encoded into a condition text embedding using a pretrained text encoder (Oquab et al. 2024). The embedding guides the dynamic gate to assign soft weights over tailors, following a mixture-of-experts style (Zhou et al. 2022; Riquelme et al. 2021) to facilitate modular reuse and enhance zero-shot generalization to novel conditions. To enhance convergence and semantic alignment during knowledge diversion, we incorporate a representation alignment module (Yu et al. 2024), which aligns condition image embeddings with shallow diffusion features.This auxiliary supervision improves early feature learning, enhances knowledge decomposition, and strengthens alignment between generated images and target conditions.

DivControl is pretrained via knowledge diversion on Subject200K (Tan et al. 2024) using 8 base conditions and evaluated on COCO2017 (Lin et al. 2014) with 10 additional unseen conditions to assess generalization and transferability. Remarkably, DivControl requires only 450K training steps ($\sim$165 GPU hours), achieving a 36.4× reduction in computational cost compared to CtrLoRA (6000 GPU hours (Xu et al. 2024)), while improving unified controllability with average CLIP-I gains of 0.05, respectively, on base conditions. For unseen conditions, DivControl demonstrates strong zero-shot generalization, generating high-quality outputs on low-shift conditions (e.g., GRAYSCALE and LINEART) without finetuning. For high-shift conditions, DivControl achieves state-of-the-art performance by finetuning only the tailors at minimal cost ($\sim$0.23 GPU hours and 200 images), offering a sharp contrast to ControlNet, which requires over 600 GPU hours per condition for retraining.

Our main contributions are as follows: 1) We introduce DivControl, the first decomposable framework for controllable image generation. By factorizing ControlNet through knowledge diversion, DivControl enables modular, interpretable, and transferable generation with substantially reduced computational overhead. 2) We propose a representation alignment mechanism that bridges condition inputs and diffusion features, enhancing controllability and accelerating convergence in controllable image generation. 3) We construct a benchmark with 18 diverse control conditions to evaluate unified controllable generation, as well as transferability and generalization of trained models. Extensive experiments demonstrate that DivControl consistently outperforms prior methods across both seen and unseen conditions.

2 Related Work

2.1 Controllable Image Generation

Diffusion models have achieved significant progress in text-to-image (T2I) synthesis (Balaji et al. 2022; Nichol et al. 2022; Peebles and Xie 2023). Recent efforts extend control to spatial conditions, such as depth and segmentation maps (Chen, Luo, and Xie 2024; Mou et al. 2024). ControlNet (Zhang, Rao, and Agrawala 2023) exemplifies this by introducing separate branches, but training separate models for each condition is computationally expensive.

To enhance flexibility, recent methods explore unified controllable generation, aiming to use a single model across diverse conditions (Qin et al. 2023; Zhao et al. 2023; Feng et al. 2025e). However, these models struggle to generalize to novel or semantically divergent conditions. We address this by introducing a unified framework that integrates modular parameter decomposition with dynamic routing. By disentangling condition-agnostic and condition-specific knowledge during training, our method enables scalable generation across diverse conditions and efficient adaptation to unseen conditions, even under limited resources.

2.2 Learngene and Knowledge Diversion

The LEARNGENE framework, inspired by biological inheritance (Feng et al. 2025a; Wang et al. 2023), encodes task-agnostic knowledge into modular neural units for efficient transfer (Feng et al. 2025b; Li, Qi, and Geng 2025; Xie et al. 2024). Existing approaches mainly emphasize knowledge compression and reuse—either through heuristic layer selection (Wang et al. 2022, 2023) or structured operations such as Kronecker products (Feng et al. 2025c). However, they largely focus on representation reuse without explicit task-level disentanglement, limiting adaptability in multi-task and cross-domain scenarios.

To address this, knowledge diversion (Xie et al. 2025) decomposes model parameters into task-agnostic learngenes and task-specific tailors, facilitating modular reuse through gated routing. We extend this framework to controllable image generation by disentangling ControlNet parameters across conditions, enabling unified multi-condition generation and efficient adaptation to novel conditions.

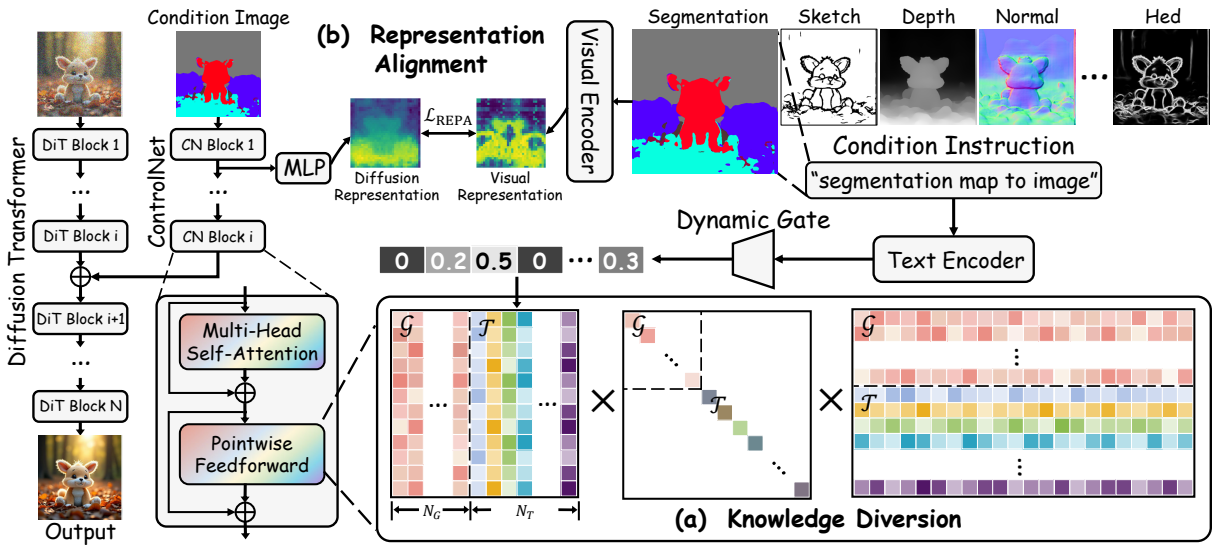

Figure 2: **Overview of the DivControl.** (a) Each weight matrix in ControlNet is factorized via SVD into condition-agnostic learngenes and condition-specific tailors. A dynamic gate routes each input to relevant tailors while jointly updating shared learngenes, enabling modular and disentangled representation across conditions. (b) Shallow features in ControlNet are aligned with condition semantics extracted by a pre-trained vision encoder, improving consistency and accelerating convergence.

3 Methods

3.1 Preliminary

Latent Diffusion Models. Latent diffusion models (LDMs) shift the generative process from high-dimensional pixel space to a lower-dimensional latent space. To enable this, an autoencoder $\mathcal{E}$ encodes an image x into a latent representation $z = \mathcal{E}(x)$, and a diffusion model is trained to reconstruct z through a denoising process, minimizing:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z,c,\varepsilon,t}\left[||\varepsilon - \varepsilon_\theta(z_t \mid c, t)||_2^2\right], \tag{1}$$

where ε_θ is a noise prediction network that predicts the noise ε added to z_t at timestep t, conditioned on c.

Conditional Generation In conditional LDMs, the condition c is derived from external modalities such as text, class labels, or condition images. We focus on image-to-image (I2I) generation, where a condition image x_{cond} provides structural guidance. A condition encoder $\mathcal{F}$ (e.g., ControlNet (Zhang, Rao, and Agrawala 2023)) maps x_{cond} to an embedding $c = \mathcal{F}(x_{\text{cond}})$, which is injected into the denoising network via adding or similar fusion strategies.

3.2 Knowledge Diversion in ControlNet

Decomposition of ControlNet We adopt a Diffusion Transformer (DiT) (Chen, Luo, and Xie 2024; Peebles and Xie 2023; Esser et al. 2024) as the backbone for generation, with the corresponding ControlNet $\mathcal{F}$ sharing the same transformer-based architecture, as shown in Figure 2. To enable flexible adaptation to diverse conditions, we apply knowledge diversion within $\mathcal{F}$, a transformer-based module composed of L layers, each containing repeated projection matrices in attention and MLP blocks. The parameter set is denoted as $\theta = \{W_q^{(1\sim L)}, W_k^{(1\sim L)}, W_v^{(1\sim L)}, W_o^{(1\sim L)}, W_{in}^{(1\sim L)}, W_{out}^{(1\sim L)}\}$[1].

Following the decomposition strategy of KIND (Xie et al. 2025), each weight matrix $W_\star^{(l)}$ ($\star \in \mathcal{S} = \{q, k, v, o, in, out\}$ and $l \in [1, L]$) is factorized via SVD as:

$$W_\star^{(l)} = U_\star^{(l)} \Sigma_\star^{(l)} {V_\star^{(l)}}^\top = \sum_{i=1}^{r} u_\star^{(l,i)} \sigma_\star^{(l,i)} v_\star^{(l,i)} \tag{2}$$

where $\Theta_\star^{(l,i)} = (u_\star^{(l,i)}, \sigma_\star^{(l,i)}, v_\star^{(l,i)})$ denotes the i-th rank-1 component, and r is the rank of $W_\star^{(l)}$. Each component $\Theta_\star^{(l,i)}$ captures a modular unit of structured knowledge.

Dynamic Diversion via Condition Text Embedding To support flexible adaptation across heterogeneous conditions, these components are explicitly partitioned into N_G condition-agnostic ***learngenes*** $\mathcal{G}$ and N_T condition-specific ***tailors*** $\mathcal{T}$, where $r = N_G + N_T$. Formally:

$$\mathcal{G} = \left\{\Theta_\star^{(l,i)} \mid i \in [0, N_G),\ \star \in \mathcal{S},\ l \in [1, L]\right\}$$

$$\mathcal{T} = \left\{\Theta_\star^{(l,i)} \mid i \in [N_G, N_G + N_T),\ \star \in \mathcal{S},\ l \in [1, L]\right\}$$

Unlike KIND (Xie et al. 2025), which relies on discrete class labels for component assignment, or CtrLoRA (Xu et al. 2024), which employs fixed task-specific adapters, we introduce a continuous diversion mechanism to enable generalization to heterogeneous and unseen conditions. To compensate for the limited semantics in condition images, each condition is paired with a manually defined condition instruction t_{cond}, which is encoded into the semantic embedding $e_{\text{txt}} = E_{\text{txt}}(t_{\text{cond}})$ using a pretrained text encoder.

[1] $W_q^{(1\sim L)}$ denotes the set $\{W_q^{(1)}, W_q^{(2)}, \ldots, W_q^{(L)}\}$. Similar notations throughout the paper follow this convention.

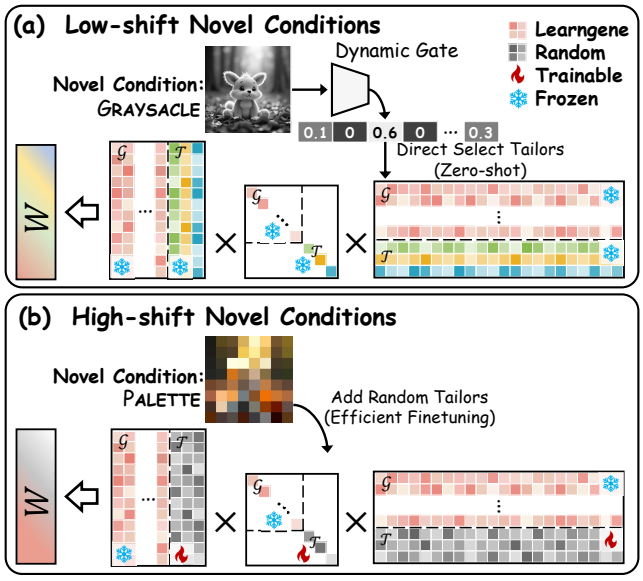

Figure 3: **Generalization to Novel Control Conditions.** (a) For low-shift conditions, DivControl leverages instruction embeddings to dynamically activate semantically aligned tailors, enabling zero-shot generation. (b) For high-shift conditions, it reuses condition-agnostic learngenes while introducing randomly initialized tailors, supporting efficient few-shot adaptation.

To enable dynamic adaptation, the condition embedding e_{txt} is processed by a lightweight gating module G, analogous to the router in Mixture-of-Experts architectures (Zhou et al. 2022; Riquelme et al. 2021). This dynamic gate produces soft weights over N_T tailor components:

$$\alpha = \text{softmax}(G(e_{\text{txt}})) \in \mathbb{R}^{N_T} \tag{3}$$

where α denotes globally shared mixing coefficients, applied uniformly across all layers $l \in [1, L]$ and projection types $\star \in \mathcal{S}$. These coefficients modulate tailor components through weighted aggregation, enabling flexible and condition-aware adaptation.

During training, the condition embedding e_{cond} activates relevant tailor components via a dynamic gate, while shared learngenes are updated across all conditions. The condition-adaptive weight matrix is constructed as a gated combination of learngenes and tailors:

$$\widetilde{W}_{\star}^{(l)} = \mathcal{G}_{\star}^{(l)} + \sum_{k=1}^{K} \alpha \cdot \mathcal{T}_{k,\star}^{(l)}. \tag{4}$$

This facilitates explicit disentanglement of condition-agnostic and condition-specific knowledge, with all components—learngenes, tailors, and the dynamic gate—jointly optimized via end-to-end training.

3.3 Representation Alignment

Representation Alignment (REPA) (Yu et al. 2024) was originally proposed to improve training efficiency and synthesis quality in class-conditioned diffusion models by aligning intermediate features with those from pretrained vision encoders (Wu et al. 2025; Tian et al. 2025; Jiang et al. 2025; Leng et al. 2025). We extend REPA to controllable image generation, where it facilitates faster optimization of the condition encoder and improves control fidelity.

Given a condition image x_{cond}, we extract a semantic embedding $e_{\text{img}} = E_{\text{img}}(x_{\text{cond}}) \in \mathbb{R}^{N \times d}$ using a frozen vision encoder. Simultaneously, a shallow feature $f_{\text{cond}} \in \mathbb{R}^{N \times d'}$ is obtained from early layers of the ControlNet $\mathcal{F}$. A lightweight MLP head $\mathcal{A}(\cdot)$ is trained to align f_{cond} with e_{img} via

$$\mathcal{L}_{\text{REPA}} = -\mathbb{E}_{z,c,\varepsilon,t}\left[\frac{1}{N}\sum_{n=1}^{N} \text{sim}(\mathcal{A}(f_{\text{cond}})^{[n]}, e_{\text{img}}^{[n]})\right] \tag{5}$$

where n is the patch index. This alignment encourages $\mathcal{F}$ to learn semantically grounded features, improving convergence while enhancing the reliability of condition-guided generation. The final objective combines denoising and alignment losses to optimize both generative fidelity and control semantics:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda \cdot \mathcal{L}_{\text{REPA}}, \tag{6}$$

where $\mathcal{L}_{\text{diff}}$ is the standard denoising loss (Eq. (1)), and λ balances the contribution of the alignment regularization.

3.4 Efficient Adaptation to Novel Conditions

Through knowledge diversion, DivControl factorizes ControlNet into condition-agnostic learngenes and condition-specific tailors, enabling rapid adaptation to novel conditions via reusable general features and modular activation of specialized components.

Given a low-shift novel condition x_{low}, we directly encode corresponding condition instruction via a pretrained text encoder to obtain $e_{\text{low}} = E_{\text{txt}}(t_{\text{low}})$, which is fed into the dynamic gate G to compute soft routing weights over N_T tailor components, as shown in Figure 3a.

$$\alpha_{\text{low}} = \text{softmax}(G(e_{\text{low}})). \tag{7}$$

Leveraging the semantic generalization of E_{txt}, this mechanism enables direct zero-shot generation for unseen conditions without gradient updates or task-specific retraining.

For conditions x_{high} with substantial semantic shifts (Figure 3b), adaptation remains lightweight by introducing randomly initialized tailors while keeping transferred parameters frozen, enabling localized fine-tuning without disrupting previously acquired general knowledge.

This modular adaptation strategy promotes generalization, facilitates cross-condition transfer, and ensures efficient adaptation with minimal overhead.

4 Experimental Setup

Dataset We perform knowledge diversion on Subject200K (Tan et al. 2024) and annotate each image with 8 basic conditions for unified conditional generation. For evaluation, we follow CtrLoRA (Xu et al. 2024) and use the COCO2017 (Lin et al. 2014) validation set, extended with 10 additional conditions spanning both low- and high-shift distributions to assess DivControl's zero-shot generalization and efficient finetuning capabilities.

	BBox			Canny			Depth			Hed			*Cost*	
	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	GPU Hour	Para.
Pixart-δ	0.255	0.766	89.43	0.283	0.473	93.69	0.232	0.823	93.20	0.246*	0.675*	95.00*	8×36 h	8×295M
UniControl	0.229*	0.777	90.11*	**0.249**	**0.500**	**94.97**	0.229	0.837*	93.68	**0.228**	**0.704**	**95.44**	5000 h	**374M**
CtrLoRA-SD	**0.221**	**0.788**	**90.50**	0.316	0.404	93.66	**0.218**	**0.841**	**94.25**	0.285	0.609	94.27	6000 h	656M
CtrLoRA-PA	0.249	0.768	89.62	0.284	0.473*	93.59	0.236	0.812	93.36	0.270	0.638	94.30	**165 h**	519M
DivControl	0.237	0.779*	89.99	0.274*	0.466	94.03*	0.223*	0.833	93.79*	0.259	0.657	94.55	**165 h**	477M*

	Sketch			Normal			Outpainting			Segmentation			*Average*		
	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I
Pixart-δ	0.260	0.718	91.70	0.400	0.705	89.96	0.061	0.898	93.42	0.458	0.652	87.92	0.274	0.714	91.79
UniControl	0.344	0.628	87.22	0.382*	**0.728**	90.58*	0.066	0.909	93.17	0.461	0.647	87.61	0.273	0.716*	91.72
CtrLoRA-SD	0.256*	0.721	92.08*	**0.377**	0.710*	**91.15**	0.072	0.893	92.95	**0.437**	**0.662**	**89.03**	0.273*	0.703	92.24*
CtrLoRA-PA	0.257	0.721*	91.67	0.400	0.697	90.16	0.058*	0.914*	93.80*	0.447	0.656*	88.35	0.275	0.710	91.85
DivControl	**0.242**	**0.741**	**92.25**	0.386	0.710	90.53	**0.053**	**0.920**	**94.51**	0.445*	0.656	88.66*	**0.265**	**0.720**	**92.29**

Table 1: **Performance on Basic Control Conditions.** We report LPIPS (↓), SSIM (↑), and CLIP-I (↑) across eight ***basic control conditions*** to evaluate generation fidelity and semantic alignment. "GPU Hour" indicates total training time, and "Para." denotes the average number of trainable parameters, reflecting overall training efficiency.

Figure 4: **Zero-shot Generalization on Low-shift Novel Conditions.** We visualize qualitative results on unseen conditions that are semantically aligned with training conditions. DivControl achieves effective zero-shot generation by leveraging knowledge diversion and dynamic gating, which route condition inputs to semantically relevant tailor components without any fine-tuning.

Basic Setting We build on PixArt-δ (Chen, Luo, and Xie 2024), adopting a DiT backbone with a 64×64 latent resolution. The ControlNet $\mathcal{F}$ shares the same 13-layer transformer architecture. Models are trained for 450K steps using AdamW with a learning rate of 1.25×10^{-5}, weight decay of 3×10^{-2}, and batch size 16. For knowledge diversion, we set the number of learngenes N_G and tailors N_T to 576, with 288 active tailor components per condition. For REPA, features from the 4-th layer of $\mathcal{F}$ are aligned with DINOv2-B (Oquab et al. 2024) embeddings via an alignment loss weighted by $\lambda = 0.05$.

5 Results

5.1 Universal Generation on Basic Conditions

DivControl enables unified controllable generation by decomposing ControlNet into reusable learngenes and condition-specific tailors through knowledge diversion, supporting scalable multi-condition control. As shown in Table 1, DivControl outperforms CtrLoRA (Xu et al. 2024) on all metrics in the average of eight basic conditions (Section 4), while reducing training time to just 165 GPU hours—much lower than the 6000 and 5000 hours required by CtrLoRA and UniControl, respectively—demonstrating both efficiency and strong generalization.

In contrast, existing methods face structural bottlenecks. UniControl (Qin et al. 2023) employs a shared encoder that lacks task-specific specialization. CtrLoRA introduces modularity through LoRA but relies on static binary gate without semantic-aware routing or representation disentanglement, limiting transferability and reuse.

DivControl mitigates these limitations through a dynamic gate, which enables semantic-aware selection of condition-specific tailors based on condition embeddings. By explicitly disentangling condition-agnostic learngenes from condition-specific modulations, it promotes modular reuse, accelerates convergence, and enhances scalability—highlighting the value of structured decomposition for unified controllable generation.

5.2 Generalization to Novel Control Conditions

We evaluate DivControl's generalization to unseen conditions, categorized as: (1) *low-shift* conditions that are semantically close to training modalities, and (2) *high-shift* conditions with substantial domain or modality gaps.

Leveraging knowledge diversion and dynamic gate, Di-

	Blur			Brush			Grayscale			Inpainting			***Cost***	
	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	GPU Time	Para.
Pixart-δ	0.179	0.923	94.70	0.168	0.769	94.56	0.209	0.708	96.92	0.149*	0.771	96.45	1.01	295M
Pixart-δ-canny	0.162*	0.933*	95.20*	0.148*	0.792*	95.65*	0.195*	0.723*	97.18*	0.160	0.771*	96.46	1.01	295M
CtrLoRA-SD	0.204	0.875	93.79	0.167	0.729	95.15	0.262	0.599	95.49	0.197	0.697	95.34	0.93	37M
CtrLoRA-PA	0.171	0.927	94.95	0.158	0.777	95.60	0.201	0.703	97.13	0.160	0.767	96.49*	**0.23**	**14M**
DivControl	**0.143**	**0.946**	**95.84**	**0.136**	**0.800**	**96.28**	**0.195**	**0.724**	**97.33**	**0.144**	**0.778**	**96.97**	**0.23**	**14M**

	Jpeg			Palette			Pixel			Shuffle			***Average***		
	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I	LPIPS	SSIM	CLIP-I
Pixart-δ	0.308	0.630	92.18	0.254	0.675	88.54	0.390	0.576	87.42	0.685	0.215	**84.80**	0.293	0.658	91.95
Pixart-δ-canny	0.300*	0.656*	92.92*	0.234	0.724*	88.89	0.366*	0.616*	89.34*	0.679*	0.229	84.48	0.281*	0.681*	92.51*
CtrLoRA-SD	0.356	0.570	92.07	0.285	0.663	87.48	0.465	0.500	86.37	0.692	0.196	84.74*	0.328	0.603	91.30
CtrLoRA-PA	0.328	0.635	92.67	0.228*	0.721	89.24*	0.383	0.588	88.76	0.681	0.237*	83.84	0.289	0.669	92.34
DivControl	**0.297**	**0.668**	**94.23**	**0.224**	**0.736**	**89.30**	**0.365**	**0.640**	**89.84**	**0.674**	**0.244**	84.74	**0.272**	**0.692**	**93.07**

Table 2: **Performance on Novel Control Conditions.** We report LPIPS (↓), SSIM (↑), and CLIP-I (↑) across 6 ***high-shift*** novel conditions and 2 low-shift novel conditions to evaluate generation fidelity and semantic alignment. "GPU Hour" indicates total training time, and "Para." denotes the average number of trainable parameters.

vControl enables zero-shot generation in low-shift settings and lightweight adaptation in high-shift scenarios. We present analyses for both cases below.

Zero-shot Generalization on Low-shift Conditions For novel conditions with minor semantic or structural deviations from training conditions, DivControl enables zero-shot generalization by embedding condition instructions into task representations that guide the dynamic gate to softly activate semantically aligned tailors. As shown in Figure 4, DivControl consistently produces high-fidelity, condition-aligned images across all low-shift novel conditions without requiring gradient updates.

In contrast, CtrLoRA (Xu et al. 2024) only assigns base ControlNet to each new condition, lacking a mechanism for semantic transfer and thus failing at zero-shot adaptation. UniControl (Qin et al. 2023) manually composes base-condition combinations for novel tasks, limiting flexibility and scalability. DivControl overcomes these limitations through input-conditioned routing over dynamically selected tailors, enabling modular reuse and scalable generalization to semantically related conditions.

Efficient Finetuning on High-shift Conditions For novel conditions with substantial distributional shifts, such as Palette and Shuffle, we transfer the parameters and initialize tailors from scratch, as these cases exhibit significant visual discrepancies and demand distinct generative priors.

As shown in Table 2, despite substantial distribution shifts, DivControl achieves competitive performance after just 3K finetuning steps ($\sim$0.23 GPU hours) on 200 images, improving average CLIP-I by 1.72 points over CtrLoRA. This demonstrates its ability to rapidly adapt to novel conditions with minimal computational overhead. Furthermore, unlike full-model retraining which ignores transferability and incurs substantial cost in both training time and parameters per condition, DivControl enables rapid, lightweight adaptation while preserving fidelity and consistency.

These results highlight the scalability and flexibility of DivControl, enabled by its structural decoupling of condition-agnostic and condition-specific knowledge. This modular design supports adaptive parameter transfer for diverse downstream demands, making it well suited for open-world scenarios with evolving or unseen conditions.

5.3 Multi-Conditional Controllable Generation

DivControl enables unified controllable image generation by disentangling condition-specific knowledge into independently activatable tailors, which further supports flexible composition of multiple control conditions through selective feature reuse—enabling more challenging, fine-grained controllable generation.

By aggregating task embeddings from multiple conditions, DivControl activates the corresponding tailor modules, enabling seamless integration of complex controls such as Lineart and Palette. As shown in Figure 5, DivControl produces high-quality and semantically consistent results even under complex, combined control conditions. This demonstrates that modular routing effectively preserves the guidance from each input while allowing them to work together for coherent generation.

5.4 Ablation and Analysis

Ablation Experiments To evaluate the impact of knowledge diversion and REPA, we ablate each component individually. As shown in Table 3, introducing knowledge diversion alone yields notable gains, with LPIPS decreasing by 0.026, SSIM increasing by 0.03, and CLIP-I improving by 1.20%, highlighting the effectiveness of modularizing condition-agnostic and condition-specific knowledge for improved generalization.

Incorporating REPA on top yields additional gains, with LPIPS further reduced by 0.012 and continued improvements in SSIM and CLIP-I, confirming its effectiveness in enhancing semantic alignment. These results underscore the complementary roles of both components: knowledge diversion enables flexible adaptation, while REPA reinforces

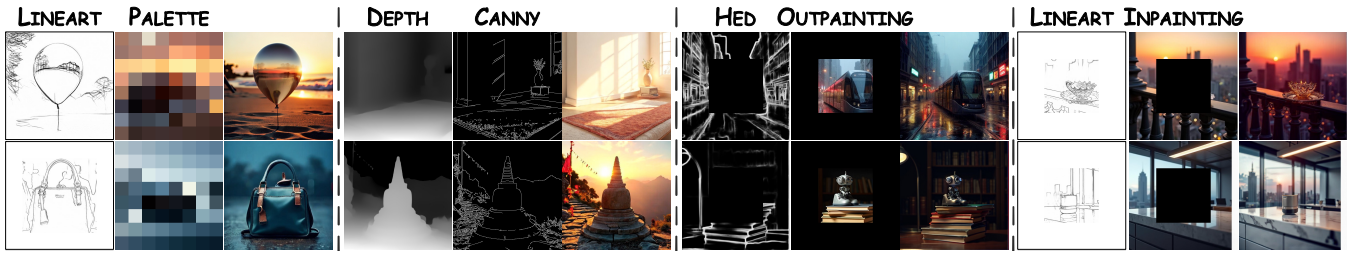

Figure 5: **Multi-Conditional Controllable Image Generation.** DivControl leverages knowledge diversion to encapsulate condition-specific knowledge into tailors, enabling flexible composition of multiple conditions. This facilitates high-fidelity, semantically aligned generation under simultaneous multi-condition guidance.

	Diversion	REPA	LPIPS↓	SSIM↑	CLIP-I↑
#1			0.328	0.663	89.43
#2	✓		0.302	0.693	90.63
DivControl	✓	✓	**0.290**	**0.696**	**91.31**

Table 3: Ablation study on DivControl.

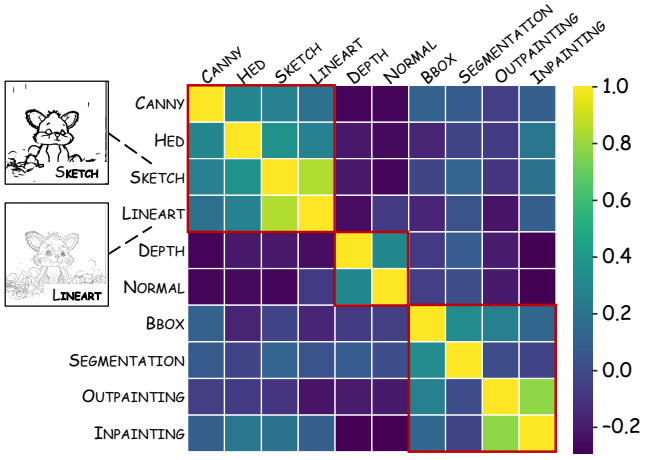

Figure 6: Inter-condition similarity derived from dynamic gate activations (α in Eq. (3)), illustrating the gate's capacity to capture semantic relationships among conditions.

semantic grounding—together facilitating high-quality, controllable generation across diverse conditions.

Analysis on Dynamic Gate To analyze the routing behavior of the dynamic gate, we visualize activation patterns across conditions. As shown in Figure 6, semantically related conditions (e.g., SKETCH and LINEART) yield similar tailor activations, whereas dissimilar ones (e.g., SKETCH vs. OUTPAINTING) exhibit divergent routing.

These results indicate that the dynamic gate effectively captures semantic similarities between conditions and modulates tailor activation accordingly. This behavior underpins DivControl's strong zero-shot performance by enabling knowledge reuse for semantically related yet unseen conditions without retraining. Moreover, the emergence of such alignment from condition instructions alone validates their

Depth	Weight	LPIPS↓	SSIM↑	CLIP-I↑	MSE↓	FID↓	FDD↓
2	0.2	0.301	0.685	90.73	40.22	10.44	0.046
6	0.2	0.303	0.682	90.70	40.34	10.44	0.046
4	0.005	0.307	0.681	90.61	40.47	10.73	0.046
4	0.5	0.315	0.672	89.98	40.76	11.01	0.053
4	0.05	**0.290**	**0.696**	**91.31**	**39.84**	**9.73**	**0.040**

Table 4: Impact of alignment depth and weight (i.e., λ in Eq. (6)) in REPA on convergence speed and stability.

effectiveness as condition representations and underscores the semantic sensitivity of the gating mechanism.

Analysis on Position and Weight of REPA We analyze the impact of REPA's alignment depth and loss weight λ on training convergence. As shown in Table 4, applying REPA at an intermediate depth (layer 4) yields the fastest convergence by balancing early-stage semantic guidance and late-stage task-specific adaptation. In contrast, shallow alignment lacks semantic expressiveness, while deeper alignment introduces condition-specific noise that impairs stability.

Additionally, moderate alignment strength ($\lambda = 0.05$) provides effective regularization without overly constraining feature learning. Excessive weighting hampers adaptability, while insufficient weighting weakens alignment signals. These results highlight the importance of carefully configuring REPA to ensure efficient and stable optimization.

6 Conclusion

In this work, we introduce DivControl, a novel training framework for ControlNet that performs knowledge diversion to construct a modular architecture during pretraining. By disentangling condition-agnostic knowledge into reusable learngenes and encoding condition-specific knowledge into lightweight tailors, DivControl supports dynamic, condition-aware assembly across diverse conditions. DivControl supports zero-shot generalization to semantically related conditions via a dynamic gate, and allows efficient adaptation to large distribution shifts through plug-in randomly initialized tailors. Extensive experiments demonstrate that DivControl consistently outperforms existing methods while substantially reducing computational overhead.

Acknowledgements

We sincerely appreciate Freepik for contributing to the figure design. This research was supported by the Jiangsu Science Foundation (BG2024036, BK20243012), the National Natural Science Foundation of China (62125602, U24A20324, 92464301, 62306073), China Postdoctoral Science Foundation (2022M720028, 2025T180432), the Xplorer Prize, and the Fundamental Research Funds for the Central Universities (2242025K30024).

References

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; et al. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Chen, J.; Luo, S.; and Xie, E. 2024. PIXART-δ: Fast and Controllable Image Generation with Latent Consistency Models. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of International Conference on Machine Learning (ICML'24)*, 1–13.

Feng, F.; Wang, J.; Yang, X.; and Geng, X. 2025a. Learngene: Inheritable "genes" in intelligent agents. *Artificial Intelligence*, 104421.

Feng, F.; Xie, Y.; Shi, R.; Shen, J.; Wang, J.; and Geng, X. 2025b. ECO: Evolving Core Knowledge for Efficient Transfer. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'25)*.

Feng, F.; Xie, Y.; Wang, J.; and Geng, X. 2025c. WAVE: Weight Template for Adaptive Initialization of Variable-sized Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'25)*, 1–10.

Feng, F.; Xie, Y.; Yang, X.; Wang, J.; and Geng, X. 2025d. Redefining <Creative> in Dictionary: Towards an Enhanced Semantic Understanding of Creative Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'25)*.

Feng, Z.; Guo, Q.; Xiao, X.; Xu, R.; Yang, M.; and Zhang, S. 2025e. Unified Video Generation via Next-Set Prediction in Continuous Domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'25)*, 19427–19438.

Feng, Z.; and Zhang, S. 2023. Efficient vision transformer via token merger. *IEEE Transactions on Image Processing*, 32: 4156–4169.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR'22)*, 1–13.

Jiang, D.; Wang, M.; Li, L.; Zhang, L.; Wang, H.; Wei, W.; Dai, G.; Zhang, Y.; and Wang, J. 2025. No Other Representation Component Is Needed: Diffusion Transformers Can Provide Representation Guidance by Themselves. *arXiv preprint arXiv:2505.02831*.

Leng, X.; Singh, J.; Hou, Y.; Xing, Z.; Xie, S.; and Zheng, L. 2025. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*.

Li, L.; Qi, L.; and Geng, X. 2025. One-Shot Knowledge Transfer for Scalable Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'25)*, 668–677.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Midjourney. 2022. Midjourney.com. https://www.midjourney.com. Accessed: 2024-11-14.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. 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 (AAAI'24)*, volume 38, 4296–4304.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of International Conference on Machine Learning (ICML'22)*, 16784–16804.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, 4195–4205.

Qin, C.; Zhang, S.; Yu, N.; Feng, Y.; Yang, X.; Zhou, Y.; Wang, H.; Niebles, J. C.; Xiong, C.; Savarese, S.; et al. 2023. UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'23)*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'23)*, 8583–8595.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*.

Tian, Y.; Chen, H.; Zheng, M.; Liang, Y.; Xu, C.; and Wang, Y. 2025. U-repa: Aligning diffusion u-nets to vits. *arXiv preprint arXiv:2503.18414*.

Wang, H.; Peng, J.; He, Q.; Yang, H.; Jin, Y.; Wu, J.; Hu, X.; Pan, Y.; Gan, Z.; Chi, M.; et al. 2025. Unicombine: Unified multi-conditional combination with diffusion transformer. *arXiv preprint arXiv:2503.09277*.

Wang, Q.; Geng, X.; Lin, S.; Xia, S.-Y.; Qi, L.; and Xu, N. 2022. Learngene: From open-world to your learning task. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'22)*, 8557–8565.

Wang, Q.; Yang, X.; Lin, S.; and Geng, X. 2023. Learngene: Inheriting Condensed Knowledge from the Ancestry Model to Descendant Models. *arXiv preprint arXiv:2305.02279*.

Wu, G.; Zhang, S.; Shi, R.; Gao, S.; Chen, Z.; Wang, L.; Chen, Z.; Gao, H.; Tang, Y.; Yang, J.; et al. 2025. Representation Entanglement for Generation: Training Diffusion Transformers Is Much Easier Than You Think. *arXiv preprint arXiv:2507.01467*.

Xie, Y.; Feng, F.; Shi, R.; Wang, J.; and Geng, X. 2024. Fine: Factorizing knowledge for initialization of variable-sized diffusion models. *arXiv preprint arXiv:2409.19289*.

Xie, Y.; Feng, F.; Shi, R.; Wang, J.; Rui, Y.; and Geng, X. 2025. Kind: Knowledge integration and diversion for training decomposable models. In *Proceedings of International Conference on Machine Learning (ICML'25)*.

Xu, Y.; He, Z.; Shan, S.; and Chen, X. 2024. CtrLoRA: An Extensible and Efficient Framework for Controllable Image Generation. *arXiv preprint arXiv:2410.09400*.

Yu, S.; Kwak, S.; Jang, H.; Jeong, J.; Huang, J.; Shin, J.; and Xie, S. 2024. Representation alignment for generation: Training diffusion transformers is easier than you think. In *Proceedings of the International Conference on Learning Representations (ICLR'25)*, 1–17.

Zavadski, D.; Feiden, J.-F.; and Rother, C. 2024. ControlNet-XS: Rethinking the Control of Text-to-Image Diffusion Models as Feedback-Control Systems. In *Proceedings of the European Conference on Computer Vision (ECCV'24)*, 343–362.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, 3836–3847.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2023. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'23)*, 11127–11150.

Zhou, Y.; Lei, T.; Liu, H.; Du, N.; Huang, Y.; Zhao, V.; Dai, A. M.; Le, Q. V.; Laudon, J.; et al. 2022. Mixture-of-experts with expert choice routing. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'23)*, 7103–7114.