



PixelPonder: Dynamic Patch Adaptation for Enhanced Multi-Conditional Text-to-Image Generation

Yanjie Pan^{1*}, Qingdong He^{2*}, Zhengkai Jiang³, Pengcheng Xu⁴, Chaoyi Wang⁵, Jinlong Peng², Haoxuan Wang¹, Yun Cao², Zhenye Gan², Mingmin Chi^{1†}, Bo Peng¹, Yabiao Wang²

¹Fudan University ²Tencent Youtu Lab ³Hong Kong University of Science and Technology

⁴Western University ⁵University of Chinese Academy of Sciences

* Equal Contribution † Corresponding author

<https://hithqd.github.io/projects/PixelPonder/>

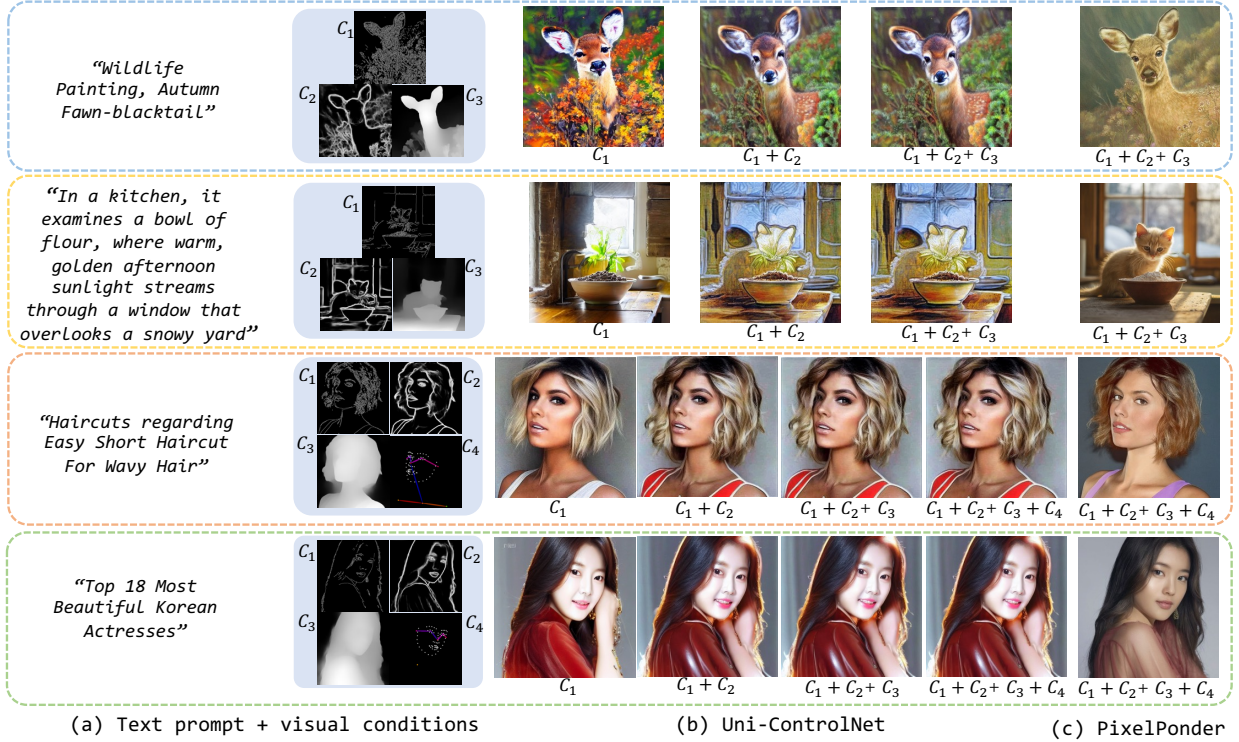


Figure 1: **Visualization comparison between Uni-ControlNet (Zhao et al. 2024) and our proposed method in different conditional controls with the same text prompt. (a, left two columns)** Text and various visual controls, where C_1 , C_2 , C_3 and C_4 denotes canny, sketch, depth and pose map respectively. **(b, middle three or four columns)** Generation results from UniControlNet. **(c, last column)** Generation results from our PixelPonder. Previous methods struggled to generate coherent results under multiple conditions, while our results maintain strong similarity to the respective visual controls.

Abstract

Recent advances in diffusion-based text-to-image generation have demonstrated promising results through visual condition control. However, existing ControlNet-like methods struggle with compositional visual conditioning - simultaneously preserving semantic fidelity across multiple heterogeneous control signals while maintaining high visual quality, where they employ separate control branches that often introduce redundant guidance during the denoising process, leading to structural distortions and artifacts in generated images. To address this issue, we present PixelPonder, a novel uni-

fied control framework, which allows for effective control of multiple visual conditions under a single control structure. Specifically, we design a patch-level adaptive condition selection mechanism that dynamically prioritizes spatially relevant control signals at the sub-region level, enabling precise local guidance without global interference. Additionally, a time-aware control injection scheme is deployed to modulate condition influence according to denoising timesteps, progressively transitioning from structural preservation to texture refinement and utilizing the control information from different categories to promote finer image generation. Extensive ex-

periments demonstrate that PixelPonder surpasses previous methods across different benchmark datasets, showing superior improvement in spatial alignment accuracy while maintaining high textual semantic consistency.

Introduction

The emergence of diffusion-based text-to-image synthesis (Rombach et al. 2022; Dhariwal and Nichol 2021; Peebles and Xie 2023; Podell et al. 2023; Ramesh et al. 2021, 2022) have driven significant progress in single-modality conditioned generation frameworks (Ye et al. 2023; Li et al. 2025; Mou et al. 2024; Cao et al. 2024; Li et al. 2023). While these approaches demonstrate enhanced controllability through visual-textual alignment, fundamental limitations persist due to modality-specific biases in visual conditioning. As evidenced in (Si et al. 2024; He et al. 2024), distinct visual control signals exhibit complementary characteristics, such as depth maps effectively regulate inter-object spatial relationships but lack fine-grained object details, whereas canny maps capture precise texture contours while disregarding global structural context. This modality specialization gap raises a critical challenge: **How to establish an effective fusion paradigm for multi-visual condition control that synergistically integrates heterogeneous visual features while resolving inter-modality conflicts?**

In practical applications, it is often necessary to describe the visual features of a significant object using multiple visual conditions to achieve precise control over its generation. Users typically desire to simultaneously control both the layout and the intricate details. However, visual conditions that are rich in layout and detail information are challenging to integrate into a single visual condition map. A common scenario involves using canny maps to represent details and depth maps to control the layout. In such case, these visual conditions are highly overlapping and complementary to one another.

Bridging the redundant information among these visual conditions and reasonably utilizing the emphasis of various visual conditions to coordinate the generation of the object is exactly the issue that this paper wants to focus on. As shown in Fig.2(a), prior efforts including Cocktail (Hu et al. 2023) and Uni-ControlNet (Zhao et al. 2024) propose to consolidate various visual conditions into unified latent representations. As illustrated in Fig.1, generating normal images under multi-modal and coupled visual condition controls remains a difficult task for these approaches. While this coarse-grained unification alleviates inter-condition conflicts through dimensionality reduction, it remains suboptimal in preserving modality-specific attributes critical for precise control (e.g., high-frequency textures in canny vs. low-frequency structures in depth). The inherent information bottleneck induced by latent compression disproportionately attenuates complementary cross-modal features, fundamentally limiting their capacity to synergize discriminative visual cues. This constitutes a critical limitation: current unification paradigms prioritize conflict mitigation over inter-modal feature reinforcement.

To address the aforementioned challenges, we introduce an innovative visual condition combination method termed

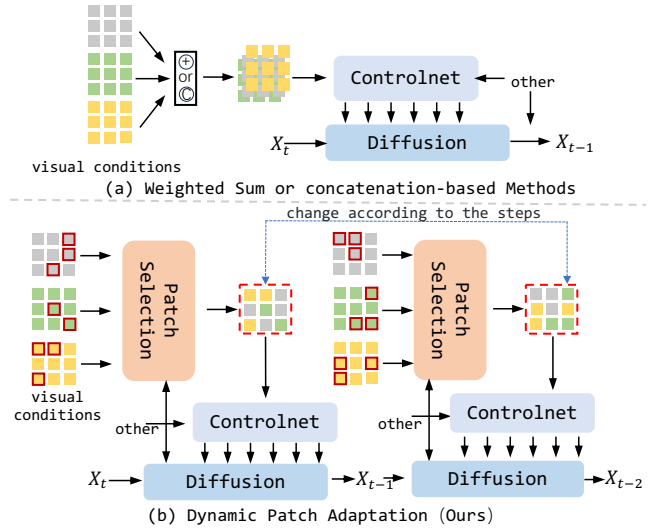


Figure 2: **Comparison of different schemes dealing with multiple condition.** (a) Existing methods integrate various visual controls into a highly consistent visual control signal in the temporal domain through concatenation or weighted summation. (b) We employ dynamic patch selection to reconstruct visual conditions at the patch level, enabling multi-visual controllable generation in the generative model.

PixelPonder. As illustrated in Fig.2(b), our method distinguishes itself from previous approaches by integrating diverse visual conditions at the patch level. Recognizing that visual conditions enriched with low-frequency or high-frequency signals significantly influence various denoising stages during the image generation process (Ho, Jain, and Abbeel 2020), we have developed an adaptive patch selection mechanism that dynamically adjusts with timesteps. This mechanism decouples each visual condition in the temporal domain, enabling the use of distinct combinations of visual conditions to regulate the denoising process at different stages. Our proposed combination method amalgamates various visual conditions with greater granularity, and facilitating complementary control of high- and low-frequency information within the temporal domain.

It is worth noting that when operating under a single visual condition control, our proposed method effectively reverts to the single visual control scheme. The efficacy of ControlNet’s single visual condition control has been extensively validated in numerous studies (Zhang, Rao, and Agrawala 2023; Zhao et al. 2024; Qin et al. 2023). Therefore, this paper does not investigate the generative capabilities of our method under single visual condition scenarios. Our main contributions are summarized as follows:

- **New Insight:** We reveal that current multi-visual condition control methods perform poorly in terms of generation quality and controllability, making it difficult for generated images to align with various visual conditions and lacking effective solutions for improvement.
- **Efficient Patch Adaptation:** We propose PixelPonder, a novel mechanism which can refine the combination of

multi-visual conditions by transitioning from the image level to the patch level, thereby enabling finer-grained controllable generation.

- **Flexible Condition Combination:** Our PixelPonder supports flexible visual control combination, addressing the redundant information conflict of multiple visual conditions for the same object.
- **Promising Results:** We provide a consolidated and public evaluation of controllability and fidelity under various conditional controls, and demonstrate that PixelPonder comprehensively outperforms existing methods.

Related Work

Diffusion-based Generative Models

Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Nichol et al. 2021; Podell et al. 2023; Ramesh et al. 2021; Chen et al. 2023) have made remarkable strides in the field of image generation. Thanks to improvements in sampling strategies (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Song et al. 2020), as well as the application of latent diffusion methods (Rombach et al. 2022), the computational resources required for these models to generate high-quality images have been substantially reduced, demonstrating promising prospects in various applications (Cao et al. 2024; Brooks, Holynski, and Efros 2023; Hertz et al. 2022; Ramesh et al. 2022). In the domain of text-to-image generation, diffusion models (Rombach et al. 2022) combined with U-Net (Ronneberger, Fischer, and Brox 2015) have successfully integrated text control into the image generation process through cross-attention mechanisms. To further enhance generative capabilities, the Transformer architecture has been integrated into diffusion models (Peebles and Xie 2023; Chen et al. 2024; Esser et al. 2024). Building on this foundation, FLUX (Labs. 2024), which introduces flow matching objectives, has achieved state-of-the-art performance in image generation.

Controllable Text-to-Image Diffusion Models

The generation of images controlled by textual descriptions often lacks spatial control. To enhance spatial control capabilities, numerous studies (Mou et al. 2024; Zhang, Rao, and Agrawala 2023; Li et al. 2023; Huang et al. 2023; Cheng et al. 2023; Yang et al. 2023a; Zhang et al. 2023) have focused on incorporating visual control signals into the generation process. In this context, ControlNet (Zhang, Rao, and Agrawala 2023) proposed a spatial control network architecture, encoding additional visual conditions into latent representations and injecting control signals into the corresponding backbone network using zero convolution, thus achieving a spatial alignment mechanism within diffusion models. Due to its straightforward approach and effective control results, it has been widely adopted in various text-to-image diffusion models (Rombach et al. 2022; Labs. 2024). Furthermore, recent research has used prompt engineering techniques (Li et al. 2023; Yang et al. 2023a) and inter-attention constraints (Chen, Laina, and Vedaldi 2024; Xie et al. 2023) to facilitate more orderly generation.

Some studies (Qin et al. 2023; Mou et al. 2024; Wang et al. 2024; Bar-Tal et al. 2023) have addressed the differences and diversity of various visual conditions in different modalities using the mix of experts (MOE) to decouple different visual conditions into distinct feature subspaces, thus establishing a unified visual control architecture. Subsequently, other research has explored the possibility of generating multiple condition controls within a single diffusion model through simple feature aggregation methods (Hu et al. 2023; Zhao et al. 2024). However, these methods are based on the low coupling of various conditions, which limits their effectiveness in tasks requiring the control of complementary conditions.

Method

Preliminary

Diffusion Transformer (DiT) (Vaswani 2017; Peebles and Xie 2023) is applied in models such as FLUX (Labs. 2024) and Stable Diffusion 3 (Esser et al. 2024). Compared to the Latent Diffusion Model (Rombach et al. 2022), these models utilize a transformer (Ruan et al. 2023) instead of U-Net as the denoising network to iteratively process noise tokens.

A model with a DiT architecture handles noisy image tokens $X \in \mathbb{R}^{N \times d}$ and text condition tokens $Y \in \mathbb{R}^{M \times d}$, where d is the dimension of the embedding layer, N and M represent the number of tokens for images and text, respectively. Additionally, they redefine the mapping from images to noise using flow matching. In this rectified mapping, the data at time t , denoted as z_t , is transformed from the real image $x_0 \in \mathbb{R}^{C \times H \times W}$ to pure noise $N \sim \mathcal{N}(0, I)$ at the following position:

$$z_t = (1 - t)x_0 + tN, \quad (1)$$

$$\vec{V} : \mathcal{N} \rightarrow \mathcal{Z}, \quad (2)$$

where $\mathcal{N}$ represents the spatial distribution of noisy images. $\mathcal{Z}$ represents the spatial distribution of real images. $\vec{V}$ is the vector field from $\mathcal{N}$ to $\mathcal{Z}$.

In FLUX, a series of flow matching modules are included, among which the denoising backbone network consists of the Double Stream Block (DSB) and the Single Stream Block (SSB).

ControlNet (Zhang, Rao, and Agrawala 2023) is designed to inject visual control information into a denoising network, thereby guiding the iterative process of noise tokens. Specifically, ControlNet freezes the parameters Θ of the backbone network $F(\cdot; \Theta)$ and creates a trainable copy Θ_c of these parameters to incorporate zero-initialized convolutional layers Z . When applied to generative models based on the U-Net architecture, a common practice is to add the visual condition C_i to the noise x_t as input to the trainable copy, while utilizing other signals s that is consistent with the backbone network. This process can be formalized as follows:

$$y = F(x_t, s, \Theta) + Z(F(x_t + M(C_1, \dots, C_n), s, \Theta_c)), \quad (3)$$

where M serves as a unified representation of various visual condition fusion methods, n represents the category id.

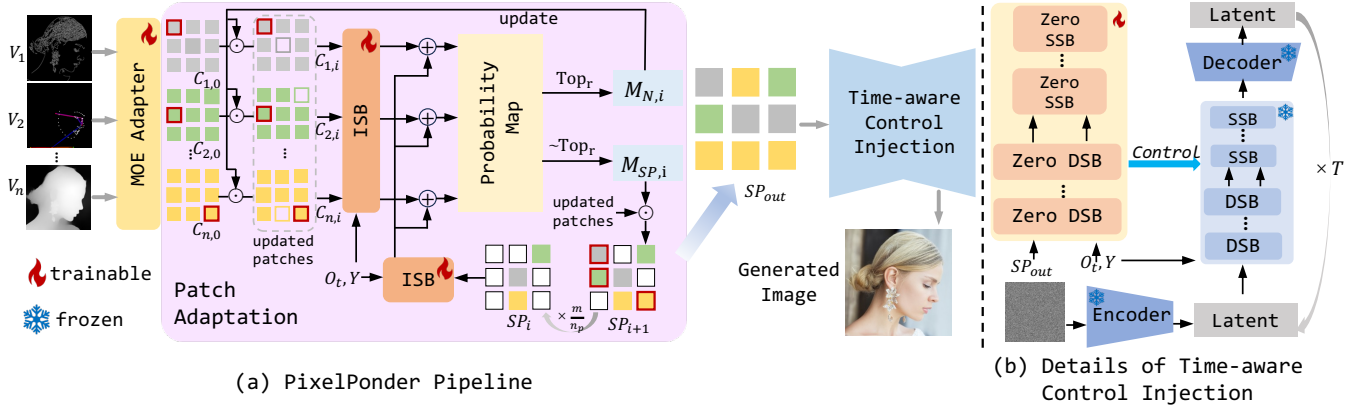


Figure 3: **Overall pipeline of the proposed.** The Patch Adaptation Module segments various visual conditions $\{V_1, V_2, \dots, V_n\}$ into patches and dynamically reorganizes the visual conditions at the patch level to obtain a unified visual control signal SP_{out} . Subsequently, SP_{out} is sent to the control network, with the Time-aware Control Injection scheme to achieve high-quality image generation under multiple visual conditions.

Patch Adaptation

The Patch Adaptation Module (PAM) aims to recompose various visual conditions at the patch level into a unified visual condition. This is achieved through an autoregressive iterative combination process that combines patches between different visual conditions. The specific algorithm flow of PAM is shown in algorithm 1.

Specifically, at time step t , the PAM initializes a zero vector unified control condition $SP_0 \in \mathbb{R}^{N \times d}$ to obtain the final unified visual control combination output SP_{out} . The combination selection is defined as an update process U_i from SP_i to SP_{i+1} , with i representing the index of the number of iterations in the update process. This iterative update process can be formalized as follows:

$$SP_{i+1} = U(SP_i, \dots, O_t), \quad (4)$$

where O_t refers to the time step information, and its specific role will be discussed in next subsection.

Assuming that the visual input to the PAM consists of N types of condition images, denoted as $V_1, V_2, \dots, V_n$, we use the corresponding visual condition feature encoders E_i to obtain various visual condition features $\mathcal{C} = \{C_1, C_2, \dots, C_n\}$, $C_k \in \mathbb{R}^{e \times d}$ and e denotes the feature dimension. To analyze and process the visual condition features in a more fine-grained manner, we can consider the visual condition feature C_k as represented by patches $P_i^m \in \mathbb{R}^d$, expressed as:

$$C_k = [P_i^0, \dots, P_i^m], \quad (5)$$

where P refers to a patch, k denotes the random category id, and m denotes the number of patches.

Algorithm 1: Patch Adaptation

Input: visual condition features $C_n \in \mathbb{R}^{N \times d}$, time step information O_t and textual feature Y .

Output: the unified visual control SP_{out}

- 1: $SP_0 = \mathbf{0}_{N \times d}$
- 2: $C_n = [P_n^0, \dots, P_n^m]$
- 3: $\mathcal{C}_0 = \{C_1, C_2, \dots, C_n\}$
- 4: $i \leftarrow 0$
- 5: $n_p \leftarrow$ number of selected patches
- 6: **while** $i \cdot n_p < m$ **do**
- 7: $M_{prob} = [\text{ISB}_C(\mathcal{C}_i, O_t, Y)] + \text{ISB}_{SP}(SP_i)$
- 8: $M_{N,i} = \text{Zero}(\text{Top}_r(M_{prob})), M_{sp,i} \sim M_{N,i}$
- 9: $\mathcal{C}_{i+1} = \mathcal{C}_i M_{N,i}$
- 10: $SP_{i+1} = SP_i + M_{sp,i} C_i$
- 11: $i \leftarrow i + 1$
- 12: **end while**
- 13: **return** SP_i

In an update process U_i , the PAM selects a set $\mathcal{P}_i = \{P_{n,i}^s\}$ composed of multiple patches $P_{n,i}^s$ from various visual characteristics C_i . For convenience, we define this selection process as F_{sp} . The composition of this set $\mathcal{P}_i$ considers two aspects of information: *a) The dominant role of textual information Y in the current time step in image generation*, and *b) The influence of previously selected patches $\mathcal{P}_0, \dots, \mathcal{P}_{i-1}$ on the overall composition of image control*. The update process from SP_i to SP_{i+1} can be more specifically expressed as:

$$U_i(SP_i, \dots, O_t) = SP_i + \mathcal{P}_i, \quad (6)$$

$$\mathcal{P}_i = F_{sp}(C_i, \mathcal{P}_0, \dots, \mathcal{P}_{i-1}, O_t, Y). \quad (7)$$

Specifically, to achieve F_{sp} , we design the Image Stream Block (ISB). To ensure that the textual control signal remains consistent globally, We modified the DSB text input selection strategy to adopt a globally unified textual feature representation Y as the input textual information. The ISB is used to obtain the weights of each patch in various visual

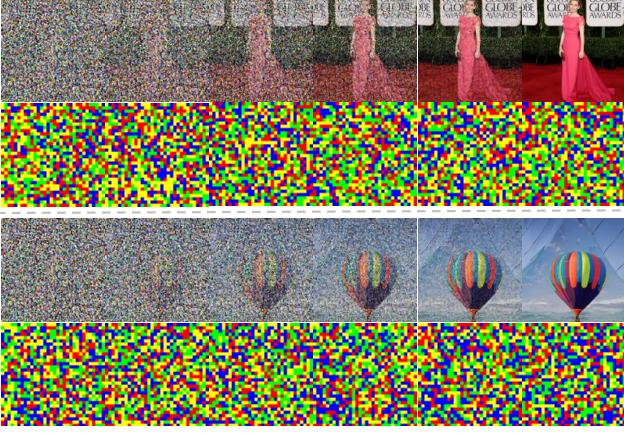


Figure 4: **Denoising results at steps 25, 20, 15, 10, 5, and 1**, where each pair includes the generated result and the selection results SP_{out} at the current time step. In each SP_{out} , canny, pose, depth and sketch are indicated in red, blue, green, and yellow respectively.

condition features under the current textual control Y at time step t , denoted as $\mathcal{W}_c = \{W_{0,i}, \dots, W_{n,i}\}$. Furthermore, to meet requirement **b**) of F_{sp} , we apply an ISB to SP_i outside of $\mathcal{C}_i$ to obtain a global weight W_{sp} that represents the global composition information of SP_i under the current U_i . The final obtained patch selection probability map M_{prob} is given by:

$$M_{prob} = [\mathcal{W}_c] + W_{sp}, \quad (8)$$

$$\mathcal{W}_c = \text{ISB}_C(\mathcal{C}_i, O_t, Y), \quad W_{sp} = \text{ISB}_{SP}(SP_i, O_t, Y). \quad (9)$$

In each iteration, to mask the influence of previously selected patch blocks $\{P_0, \dots, P_{i-1}\}$ on the combination selection process U_i , we use a binary mask $M_{k,i}$ to update C_k^{i-1} to obtain C_k^i . The binary mask $M_{k,i}$ is derived from M_{prob} :

$$M_{N,i} = \text{Zero}(\text{Top}_r(M_{prob})), \quad (10)$$

$$[\bigcup_{k=1}^N M_{k,i}] = M_{N,i}, \quad (11)$$

where Top_r retrieves the positions of the top r patches with the highest probabilities in M_{prob} , and Zero sets the corresponding positions in the binary mask to zero.

Fig.4 illustrates the visualization results of SP_{out} obtained by PAM in different time steps, as well as the generated images. In the denoising process, the generated images exhibit a noticeable tendency towards patchiness. The segmentation contours between the generated image patches are prominent. Both examples demonstrate the characteristics of PAM: a) PAM exhibits a balanced tendency in its selections over various time steps. b) For patches at the same location, the global selections include patches from all categories.

It is noted that C_n^0 is equivalent to C_n , which means that we do not process C_n in U_0 . The representation of the visual condition C_n during the iteration process is as follows:

$$C_k^i = \begin{cases} \mathbf{E}_k(V_k) & \text{if } i = 0 \\ M_{k,i} C_k^{i-1} & \text{if } i > 0. \end{cases} \quad (12)$$

While obtaining $M_{N,i}$, we can also acquire the mask $M_{sp,i}$ corresponding to the selected patch P_i during the update process U_i :

$$M_{sp,i} = \sim M_{N,i}, P_i = M_{sp,i} \mathcal{C}_i, \quad (13)$$

where $\sim$ represents the negation operation.

Time-aware Control Injection

The time-aware control injection is designed to fully utilize the visual control effects of various visual conditions at different time steps during the denoising process. This is achieved by introducing additional temporal information O_t in the PSM at each time step, which controls the patch selection tendency in each update process U . To ensure that the patch selection tendencies in all selection processes U within a time step are consistently influenced by the temporal information, we use a unified temporal signal O_t at the current time step for control, as shown in Eq. 4.

Specifically, we introduce the time-step information O_t into the visual conditions and the global ISB module. This allows the ISB to simultaneously acquire time-step control information to obtain patch selection weights that are more aligned with the denoising temporal flow, which can be formalized as follows:

$$\text{Control} = D'(SP_{out}, Y, O_t, \Theta_c), \quad (14)$$

$$I_r = D(N, \text{Control}, Y, \Theta), \quad (15)$$

where D denotes the original model, Θ indicates the model parameters, D' refers to the control network, and Θ_c signifies the model parameters of the control network, I_r represents the generated image. Considering the large number of parameters in the model, we reduce the number of modules in the control network and align the image streams sequentially to the corresponding backbone modules.

To achieve stable control of visual composition under time awareness, we initialize the last linear layer to zero from the ControlNet and FLUX, denoted as zero-initialized Double Stream Block (**Zero DSB**) and zero-initialized Single Stream Block (**Zero SSB**), which can be formulated as:

$$\text{Attention}(Q, K, V) = Z((\frac{QK^T}{\sqrt{d_k}}), V). \quad (16)$$

Training

Objective. Given a set of multi-visual prompts $\mathcal{C}$ and a text prompt Y , we fine-tune our method to reconstruct the original image x . During the training process, we freeze the backbone network and only optimize the patch selection network and the zero controllable network. Our training objective combines V and V' to obtain the visual representation $V_{n \rightarrow c}$ under the visual prompts $\mathcal{C}$. This encourages the visual control network to learn the internal visual condition control flow V' based on the V pre-learned by the backbone network. The loss function $\mathcal{L}$ is formalized as:

$$\mathcal{L} = \|(N - x_0) - \int_0^1 \vec{V} \oplus \vec{V}'(z_t, \mathcal{C}, Y, t) dt\|^2, \quad (17)$$

where the $\oplus$ represents the vector sum of two fields.

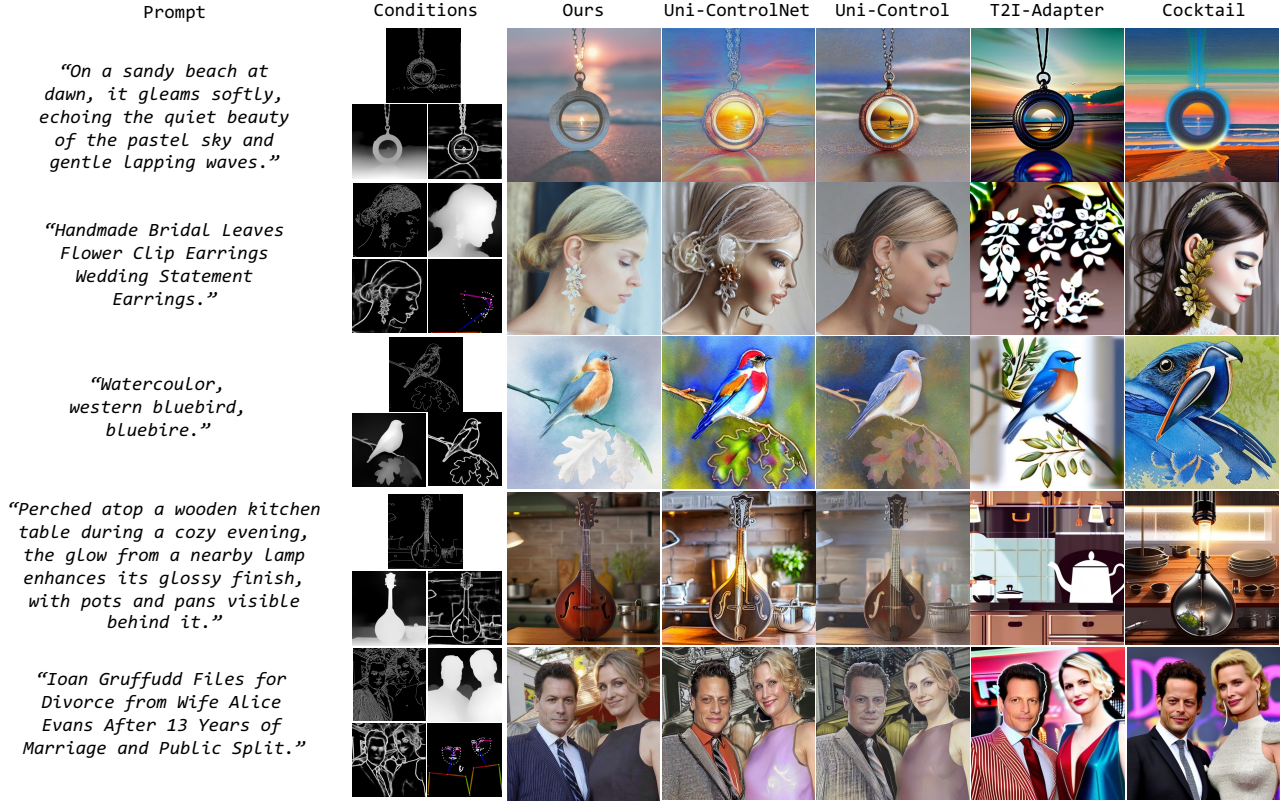


Figure 5: **Visualization comparison** between official or re-implemented methods and our proposed model in different conditional controls.

Experiments

We validate the effectiveness of PixelPonder on the MultiGen-20M (Li et al. 2025) and Subject-200K (Tan et al. 2024) datasets. Our evaluation primarily focuses on several leading methods, including T2I-Adapter (Mou et al. 2024), Uni-ControlNet (Zhao et al. 2024), UniControl (Qin et al. 2023), and Cocktail (Hu et al. 2023) in the realm of multiple controllable text-to-image diffusion models. It is worth noting that we conducted experiments on Cocktail using only pose and sketch maps, as it only supports pose, sketch, and segmentation maps. Although the models of other approaches such as AnyControl (Sun et al. 2024) are public, their code cannot be successfully run after many attempts. More implementation details can be found in the supplementary material.

Main Results

Comparison of Controllability. As shown in Tab.1, we report the controllability comparison results across different datasets, where the structural similarity index (SSIM) is employed to evaluate image consistency. Our PixelPonder surpasses other methods with improvements of 8.41% and 19.29% on two benchmark datasets, establishing a new state-of-the-art in image generation. This notable accomplishment demonstrates that PixelPonder is capable of processing intricate combinations of multiple spatial condi-

tions, generating images that are highly consistent with the given spatial constraints.

Comparison of Image Quality. Previous methods often encounter two primary issues when handling multiple visual control conditions: a) Coarsening of lines within the image, and b) Disappearance or distortion of the main subject. In contrast, PixelPonder, leveraging its unique architectural design, effectively avoids these issues by harmonizing multiple visual control conditions without compromising image quality. This is demonstrated by the quantitative results in Tab.1, including significant improvements in FID (12.82 on MultiGen-20M and 20.34 on Subject-200K) and leading MUSIQ scores, and qualitative comparison in Fig.5 further supports the effectiveness of PixelPonder in handling complex visual conditions.

Comparison of CLIP Score. Tab.1 compares CLIP Score metric across methods under maximal conditions. The results reveal a consistent trade-off: methods with fewer visual constraints achieve higher text adherence. Specifically, PixelPonder demonstrates superior performance under full conditional control, outperforming the compared methods. Although Cocktail attains higher scores under significantly fewer visual constraints, the strong text adherence of Cocktail significantly compromises the consistency of its visual performance. As illustrated in Fig.5, the images generated by Cocktail exhibit pronounced issues of object loss and mu-

Methods	Conditions	MultiGen-20M				Subject-200K				Inference Time (↓)
		FID (↓)	CLIP Score (↑)	SSIM (↑)	MUSIQ (↑)	FID (↓)	CLIP Score (↑)	SSIM (↑)	MUSIQ (↑)	
T2I-Adapter	all	66.95	71.47	24.03	57.95	64.72	74.35	31.76	55.07	9.33
Uni-ControlNet	all	32.58	78.08	29.37	65.85	44.35	77.40	37.98	66.75	7.22
UniControl	all	25.15	74.09	<u>35.58</u>	72.05	<u>30.95</u>	72.96	47.31	67.50	7.39
Cocktail	pose+hed	<u>24.67</u>	79.87	24.14	67.13	51.51	79.38	29.59	64.16	4.94
PixelPonder(Ours)	all	11.85	<u>78.60</u>	43.99	<u>69.54</u>	10.61	77.12	66.60	<u>67.32</u>	9.59

Table 1: Comparison on the MultiGen-20M (Li et al. 2025) and Subject-200K (Tan et al. 2024) datasets. The best and second-best results are highlighted in **bold** and underline, respectively. In the **Conditions** column, “all” denotes the use of canny maps, sketches, pose maps, and depth maps simultaneously during generation.



Figure 6: **Comparison of multi-condition image generation** by PixelPonder with different combinations against single-condition image generation by ControlNet++.

tations when compared to those produced by other methods. This highlights PixelPonder’s unique capability to balance precise visual control with competitive text adherence.

Comparison of Inference Speed. Tab.1 presents the inference speed. Due to the fact that Cocktail involves only half the number of conditions compared to other methods, any direct comparison of time efficiency may not be entirely representative or meaningful. Patch Adaptation introduces a marginal increase in generation time; however, given the substantial improvement in image quality, the associated computational overhead is acceptable.

Condition Types Comparison. To examine the effect of the number of condition types, Fig.6 shows the visual results under various condition combinations, with qualitative evaluation metrics provided in Tab.2. Examples in Fig.6 demonstrate that as the number of conditions increases, the layout and texture within images become increasingly refined and accurate, demonstrating that the control effects of different visual conditions are not identical. Furthermore, as can be seen from Tab.2, an increase in visual conditions is beneficial for the overall quality and controllability of the images, which aligns with the observed visual effects. As evidenced by the comparison with ControlNet++ (Li et al. 2025), integrating multi-conditions improves control precision and enables finer-grained manipulation of the generated output.

Ablation Study

Patch Selection. In Fig.7, we illustrate the statistical distribution of each visual condition patch during the denoising steps, based on tens of millions of samples. The range across all stages exceeds 1%, indicating a significant differ-

Methods	Conditions	MultiGen-20M		
		FID (↓)	SSIM (↑)	MUSIQ (↑)
Source	-	-	-	69.30
ControlNet++	canny	17.69	36.69	65.67
ControlNet++	hed	13.93	42.12	71.22
ControlNet++	depth	17.56	27.79	71.23
Ours	pose	30.00	27.91	58.16
Ours	pose+depth	19.76	31.63	62.63
Ours	pose+depth+hed	12.41	40.48	67.82
Ours	all	11.85	43.99	69.54

Table 2: **Results of combining different condition types.** A quantitative comparison with models controlled by single visual conditions and various combinations of visual conditions. “Source” refers to the original reference image.

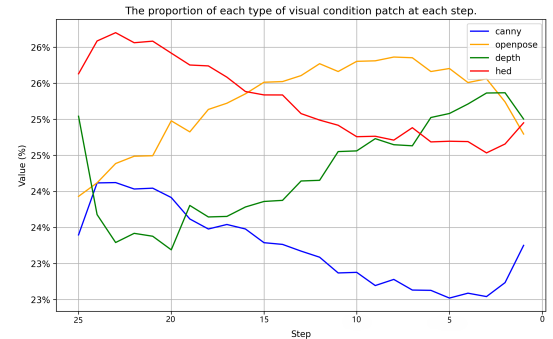


Figure 7: **Statistical distribution of each visual condition patch** during the denoising steps.

ence between the Patch Selection strategy and random sampling. Furthermore, The trends show that generating low-frequency or high-frequency regions does not require consistency with the initial layout; instead, it benefits from incorporating visual information from other frequency domains to meet diverse visual requirements. Visual comparisons are provided in the supplementary material.

Patch Size. Tab.3 presents the performance of PixelPonder under different patch sizes. The experimental results indicate that finer-grained image condition combination can better resolve conflicts among various visual elements in the visual conditions.

Zero Controllable Flow. As shown in Tab.4, the use of zero controllable flow significantly enhances the aesthetic quality and color richness of the images. However, this comes with a slight loss of feature consistency in the images, while maintaining better visual coherence.

Patch Size	MultiGen-20M			
	FID (↓)	CLIP Score (↑)	SSIM (↑)	MUSIQ (↑)
2	11.26	78.62	43.21	69.15
4	12.80	78.24	40.73	68.43
8	13.46	78.07	39.75	68.15

Table 3: **Ablation on patch size.** We tested the impact of patch size in Patch Selection on controlling the generation.

Zero	MultiGen-20M		
	FID (↓)	SSIM (↑)	MUSIQ (↑)
×	11.26	43.21	69.15
✓	11.85	43.99	69.54

Table 4: **Ablation on zero controllable flow.** × represents the non-use of zero controllable flow, while the ✓ represents the use of zero controllable flow.

Conclusion

In this paper, we present PixelPonder, a novel framework for compositional visual conditioning in diffusion-based image generation. Our key contribution lies in addressing the fundamental limitation of existing methods in handling redundant guidance from multiple controls. Specifically, we introduce two novel components, a Patch-level Adaptive Condition Adaptation mechanism that dynamically resolves spatial conflicts through learnable attention gates, and a Time-aware Control Injection scheme that orchestrates condition influence across denoising phases. Extensive experiments on several benchmarks demonstrate significant improvements over state-of-the-art methods. With PixelPonder, users can depict different aspects of objects by utilizing various visual conditions, thereby accurately realizing their creations.

References

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instruct-pix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 679–698.

Cao, P.; Zhou, F.; Song, Q.; and Yang, L. 2024. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*.

Chen, J.; Wu, Y.; Luo, S.; Xie, E.; Paul, S.; Luo, P.; Zhao, H.; and Li, Z. 2024. Pixart- $\{\delta\}$: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

Chen, M.; Laina, I.; and Vedaldi, A. 2024. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5343–5353.

Cheng, J.; Liang, X.; Shi, X.; He, T.; Xiao, T.; and Li, M. 2023. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

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 *Forty-first International Conference on Machine Learning*.

He, Q.; Peng, J.; Xu, P.; Jiang, B.; Hu, X.; Luo, D.; Liu, Y.; Wang, Y.; Wang, C.; Li, X.; et al. 2024. DynamicControl: Adaptive Condition Selection for Improved Text-to-Image Generation. *arXiv preprint arXiv:2412.03255*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, M.; Zheng, J.; Liu, D.; Zheng, C.; Wang, C.; Tao, D.; and Cham, T.-J. 2023. Cocktail: Mixing multi-modality control for text-conditional image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Huang, L.; Chen, D.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5148–5157.

Labs., B. F. 2024. Flux: Official inference repository for flux.1models Accessed:2024-11-12.

Li, M.; Yang, T.; Kuang, H.; Wu, J.; Wang, Z.; Xiao, X.; and Chen, C. 2025. ControlNet ++ : Improving Conditional Controls with Efficient Consistency Feedback. In *European Conference on Computer Vision*, 129–147. Springer.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

- 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*, volume 38, 4296–4304.
- Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.
- Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
- 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. *arXiv preprint arXiv:2305.11147*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- 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.
- Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.
- Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1623–1637.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, 234–241. Springer.
- Ruan, L.; Ma, Y.; Yang, H.; He, H.; Liu, B.; Fu, J.; Yuan, N. J.; Jin, Q.; and Guo, B. 2023. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10219–10228.
- Si, C.; Huang, Z.; Jiang, Y.; and Liu, Z. 2024. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4733–4743.
- Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Sun, Y.; Liu, Y.; Tang, Y.; Pei, W.; and Chen, K. 2024. AnyControl: Create Your Artwork with Versatile Control on Text-to-Image Generation. *arXiv preprint arXiv:2406.18958*.
- 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*.
- Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- Wang, X.; Darrell, T.; Rambhatla, S. S.; Girdhar, R.; and Misra, I. 2024. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6232–6242.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.
- Xie, J.; Li, Y.; Huang, Y.; Liu, H.; Zhang, W.; Zheng, Y.; and Shou, M. Z. 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7452–7461.
- Xie, S.; and Tu, Z. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, 1395–1403.
- Yang, Z.; Wang, J.; Gan, Z.; Li, L.; Lin, K.; Wu, C.; Duan, N.; Liu, Z.; Liu, C.; Zeng, M.; et al. 2023a. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14246–14255.
- Yang, Z.; Zeng, A.; Yuan, C.; and Li, Y. 2023b. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4210–4220.
- Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.
- 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*, 3836–3847.
- Zhang, T.; Zhang, Y.; Vineet, V.; Joshi, N.; and Wang, X. 2023. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2024. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36.



PixelPonder: Dynamic Patch Adaptation for Enhanced Multi-Conditional Text-to-Image Generation

Supplementary Material

Implementation Details

Dataset. Our training dataset is randomly sampled from the MultiGen-20M (Li et al. 2025) dataset, resulting in a training set of 2.5 million samples. Each training sample contains an input image x , a text prompt, and a set of visual conditions $\mathcal{C} = \{C_1, C_2, \dots, C_n\}$. This set $\mathcal{C}$ of visual conditions includes various types of visual cues, such as canny maps, sketches, depth maps, and pose maps, obtained from several other methods (Yang et al. 2023b; Ranftl et al. 2020; Canny 1986; Xie and Tu 2015). In the testing dataset, due to the small size of the MultiGen-20M test set, and to ensure fairness in comparison, we combine the test and validation sets of MultiGen-20M to form a test set of up to 5,500 text-image pairs. Additionally, to increase the confidence in our conclusions, we randomly sample 5,000 text-image pairs from the Subject-200K (Tan et al. 2024) dataset as a second test set. This dataset provides more complex text compared to MultiGen-20M, and the Subject-200Ks in each image are clearly defined.

Training Setting. We conduct the training on 8 GPUs with 96GB of VRAM, using a batch size of 4 for every GPU, for a total of 200,000 training steps. The AdamW optimizer is employed with a learning rate of 2×10^{-5} . The default setting for patch size is 2, and the default number of patches selected per U_i is set to $\frac{W}{\text{patch size}}$ (where W is the width of the image features). The default depth of the ISB module is 1, and the default number of Zero DSBs and Zero SSBs is set to 2 and 4, respectively. During training, we randomly select 3 out of 4 types of conditions (canny maps, depth maps, sketch maps, and pose maps) as visual control inputs. Regarding the pose condition, since not all images contain humans, we do not discard the pose condition to ensure its effective training.

Inference Setting. During the inference process, the sampling step for PixelPonder is set to 25. For the comparative experiments with other methods, such as UnicontrolNet and Cocktail, the hyperparameters during inference is set to the official default configurations. In all visual displays except for the multiple generated images of the same object, the global random seed is set to 100. It is worth mentioning that assuming the visual conditions consist of $H \times W$ tokens, the number of patches selected in each update process U_i is W , and the total number of update processes is H .

Evaluation Metrics. For multi-visual condition control tasks, we evaluate generation quality, controllability, and text-image consistency using a comprehensive set of math-

ematical and model-based metrics. **Generation quality** is assessed through **FID** (Heusel et al. 2017) and **MUSIQ** (Ke et al. 2021). FID, a model-based metric, quantifies the similarity between feature distributions of real and generated images, providing a robust measure of image fidelity. MUSIQ, another perceptual model-based metric, evaluates image quality from a human perceptual perspective by capturing both local and global quality attributes. **Controllability** is measured using **SSIM** (Wang et al. 2004), a mathematically grounded metric that computes the structural similarity between reference and generated images, offering a precise assessment of visual fidelity. The evaluation metric for **inference speed** is the time cost per image, measured in seconds per image (s/per image). Finally, **text-image consistency** is evaluated via **CLIP Score** (Hessel et al. 2021; Radford et al. 2021), a model-based metric that measures the semantic alignment between input text and generated images.

Scale Study

To investigate whether PixelPonder demonstrates better image quality with larger parameter scales and training datasets, we evaluated the fidelity and consistency of the generated images using FID, SSIM and MUSIQ on different data scales and module depths.

Data Scale. We evaluate the performance of PixelPonder on different scales of training datasets, with the results shown in Tab.5. Due to limitations in computational resources and time, we only explored up to 2.5 million data. As shown in the table, PixelPonder demonstrates significant improvements in FID and SSIM as the scale of the training dataset increases. However, MUSIQ, constrained by the original reference images, does not show further improvement after a substantial increase in the size of the data set. Compared to other methods, such as ControlNet++ and UniControl, our method achieves results comparable to those obtained by tens of millions of training samples, while being trained on only millions of samples. This indicates that there is still room for improvement in the precise control capabilities of PixelPonder.

Patch Selection Analysis

To further explore the specific internal processes of patch selection, we conduct some investigations, including the selection tendencies of patch selection for various visual condi-

Data scale	MultiGen-20M		
	FID ($\downarrow$)	SSIM ($\uparrow$)	MUSIQ ($\uparrow$)
105k	12.20	40.79	68.93
686k	11.43	41.13	69.16
2.5m	11.26	43.21	69.15

Table 5: **Ablation on data scale.** FID, SSIM and MUSIQ are reported on MultiGen-20M datasets. k represents thousand, and m represents million.

tions, as well as the characteristics of the denoising process based on patch selection.

Previous methods exhibit a progression of features from coarse to fine during the denoising process. Specifically, generative models first create the rough outlines and layouts of various objects, followed by the addition of details, which means generating from low-frequency regions to high-frequency regions. Generally, we believe that generating low-frequency visual elements requires low-frequency information, while high-frequency visual elements require high-frequency information.

Fourier	MultiGen-20M		
	FID ($\downarrow$)	SSIM ($\uparrow$)	MUSIQ ($\uparrow$)
$\times$	12.79	44.97	68.38
$\checkmark$	13.21	44.49	68.09

Table 6: **Ablation on fourier.** $\times$ represents the non-use of fourier, while the $\checkmark$ represents the use of fourier.

To further validate our idea, we design a Fourier-based low-frequency and high-frequency signal correction, aiming at adjusting the intensity of low-frequency and high-frequency signals to be consistent with previous conclusions. Its mathematical form is as follows:

$$f_{low} = \alpha M_{low} f t \quad (18)$$

$$f_{high} = \frac{\tilde{M}_{low} f}{\min(t, \frac{2}{3})} \quad (19)$$

where t is the current time step, M_{low} is a low-frequency mask, $\tilde{M}_{low}$ is a high-frequency mask, α is the correction hyperparameter, f is the spectrogram, and f_{low} and f_{high} represent the corrected low-frequency and high-frequency signals, respectively. f_{low} decreases over time steps, while f_{high} increases. We incorporate this Fourier correction into the control network, and the results are shown in Tab.6. It demonstrates that the image quality generated by the Fourier correction constructed under the previous conclusions has declined in various aspects, which indirectly supports the validity of our conclusions.

In the examples shown in Fig.4 of the main paper, due to the fine-grained division of patches, the pose maps are mostly blank noise at many patch locations, resulting in a lack of control effect. However, contrary to our expectations, in the invalid areas corresponding to the pose maps, patch

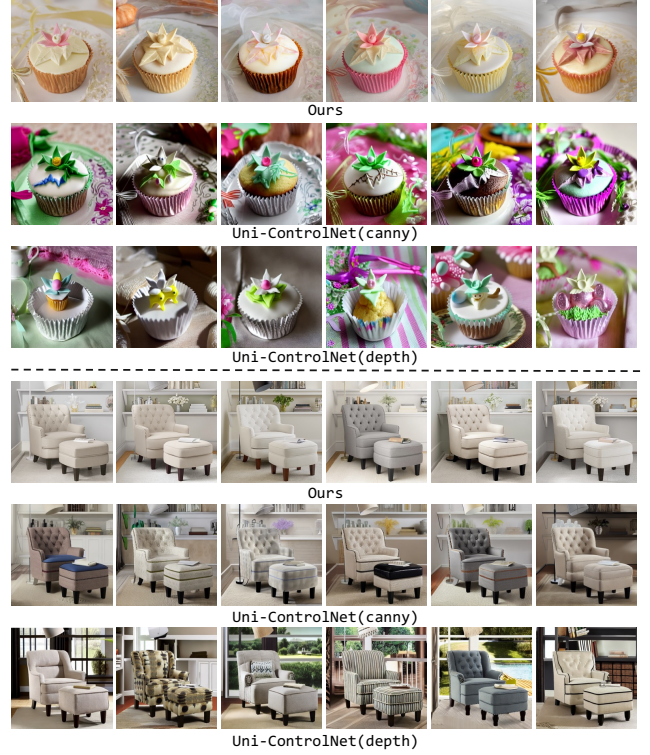


Figure 8: **Comparison of multiple random generations under single-condition and multi-condition.** *Ours* employs all available conditions, whereas *canny* and *depth* denote the use of only the canny or depth conditions, respectively.

selection still chooses the blank noise from the pose maps to perturb the generation, which leads to better generation results. This is supported by the comparisons in Tab.2 of the main paper, where the generation results under all conditional control scenarios significantly outperform other combinations of conditions, particularly in consistency evaluation metrics such as FID and SSIM. This indicates that appropriately adding noise during the denoising process does not negatively impact image generation; rather, it frees up more space for other visual condition controls to exert their influence, which is consistent with our conclusions regarding the selection tendencies of various visual conditions in patch selection.

Generation Stability Comparison

As shown in Fig.8, the previous single-condition control framework encounters significant challenges in preserving both stability and consistency regarding the detailed features and morphological characteristics of target objects across multiple generated images. Particularly in the cake and furniture scenarios, Uni-ControlNet manifests several limitations, including texture degradation, contour deformation, and background inconsistency. In stark contrast, PixelPonder demonstrates superior performance in maintaining both consistency and stability across these dimensions. A notable example can be observed in the cake scenario, where

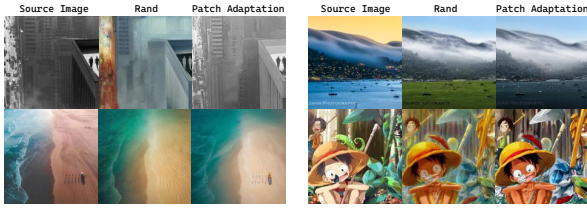


Figure 9: **Generation quality comparison** between Random Selection and Patch Adaptation. Random selection can lead to a loss of information, resulting in reduced capability to generate details and clarity.

PixelPonder effectively maintains precise control over the cream’s morphological features while simultaneously ensuring consistent dimensional attributes of the cake across all generated images.

As shown in Fig.9, randomly selected visual results exhibit significant deviations in terms of image layout, fine details, and intra-image contextual understanding of the garment-person interaction. Although the statistical difference is only a few percentage points, the visual discrepancy is highly noticeable. This demonstrates the effectiveness of the patch adaptation strategy in the generation process.

Image Consistency Exploration

To investigate whether the images generated by PixelPonder are consistent with the visual control conditions in terms of visual elements, we provide additional relevant visual demonstrations in Fig.10 and Fig.11.

In each visual example, we present the reference along with its corresponding visual conditions. Additionally, the generated images are displayed under the visual conditions of the reference, as well as the extracted visual conditions from these generated images to investigate the consistency of each method with respect to various visual conditioning details in terms of visual fidelity and adherence to the specified conditions.

More Qualitative Comparison

In Fig.10 and Fig.11, we present more visual results across various scenes. Previous methods do not separate multiple classes of visual conditions in visual control; instead, they inject multiple visual conditions simultaneously at a single time step, redundantly introducing visual elements of can-nys. Our proposed PixelPonder, through patch-level separation of visual conditions and temporal domain separation, avoids the generation of overlapping conflicting visual signals and achieves the injection of complementary information among various visual elements in the temporal domain, resulting in superior generation performance under multi-visual condition control.

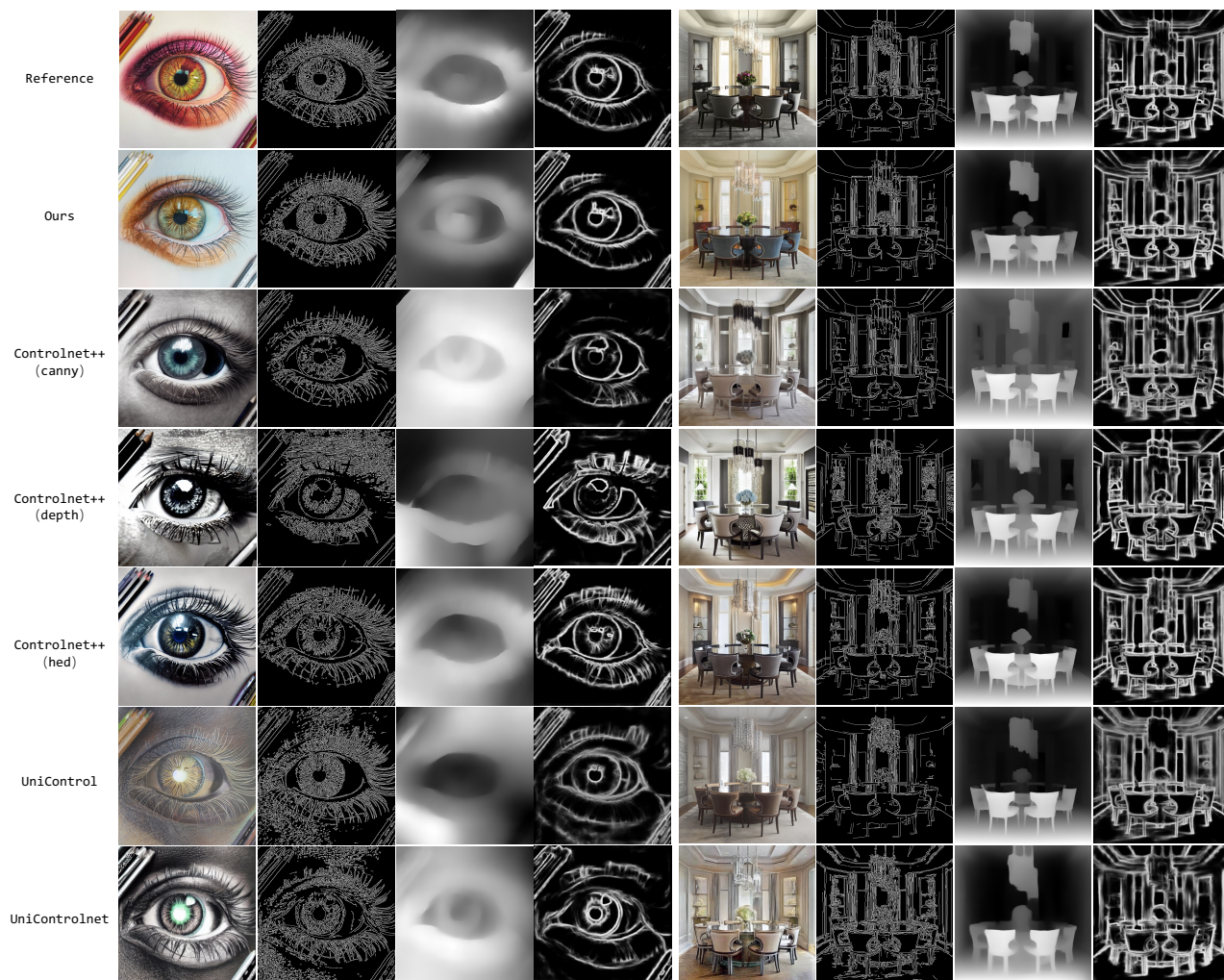


Figure 10: **Image consistency comparison** between official or re-implemented methods and our proposed model.

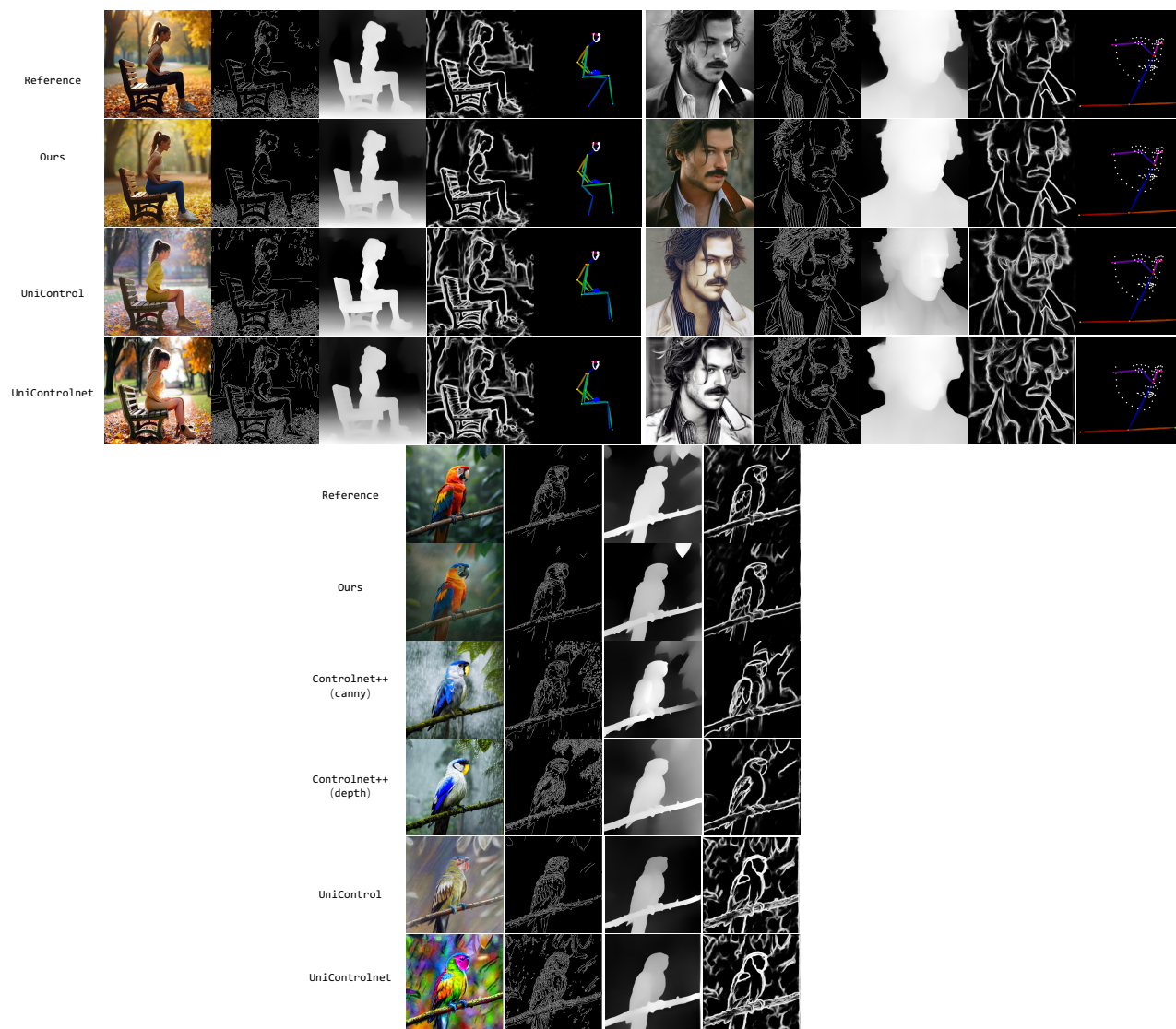


Figure 11: **Image consistency comparison** between official or re-implemented methods and our proposed model.