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

Supplementary Material

The supplementary material presents the following sections to strengthen the main manuscript:

- A. Preliminaries.
- B. Pseudo-code of DynFusion.
- C. Implementation Details.
- D. More Ablation Studies.

A1. Preliminaries

Diffusion Transformer. The condition-guided Diffusion Transformer (DiT) has gradually replaced Latent Diffusion Model (LDM) [4], which utilize UNet as denoising network for iterative denoising, and has become mainstream generative model.

DiT combines the noise tokens and text tokens at token dimension to serve as input for the denoising backbone. The size of these tokens remains consistent throughout the entire process. In FLUX, Muti-Modal Attention (MMA) [6] is the core component of the DiT block, while Rotary Position Embedding (RoPE) [7, 8] is incorporated into it to encode spatial information:

$$\tilde{X}_{i,j} = X_{i,j} \cdot R\left(i,j\right), \tag{1}$$

for query Q and key K vector, RoPE applies rotation matrices $R(i,j)$ based on tokens' position coordinates (i,j) in the 2D grid. Furthermore, MMA can be depicted as:

$$\mathrm{MMA}\left(Q,K,V\right)_{X,C_T} = \mathrm{softmax}(\frac{\tilde{Q}\tilde{K}^{\top}}{\sqrt{d}})V, \tag{2}$$

where Q, K, V are derived from the concatenation of noisy image and text tokens. MMA realizes the integration of multi-mode features. Additionally, flow matching is used to transform image to noise:

$$\overrightarrow{v}(\mathbf{z}_t, t) = \frac{\mathrm{d}\mathbf{z}_t}{\mathrm{d}t}, \tag{3}$$

$$\mathbf{z}_{t+\triangle t} = \mathbf{z}_t + \triangle t \cdot \overrightarrow{v}(\mathbf{z}_t, t), \tag{4}$$

where $t \in [0,1]$, $\mathbf{z}_t$ can be regarded as a data point at time t. $\overrightarrow{v}(\mathbf{z}_t, t)$ is a vector field that defines the magnitude and direction of the change of data points at each timestep. This vector is learned by denoising network.

Multi-Conditional Generation. To endow DiT with the function of multi-condition controllable generation, we can refer to the relevant approaches of ControlNet, and add the condition latent variable to the noise latent variable:

$$y = \mathcal{F}(C_T, X; \Theta) + \mathcal{Z}(\mathcal{F}_c(C_T, X, C_V^{1:n}; \Theta_c); \Theta_Z), \tag{5}$$

where $F(\cdot;\Theta)$ represents the frozen backbone, $F_c(\cdot;\Theta_c)$ serves as various condition fusion methods, *e.g.* sequential control adapters or shared control adapters, and $Z(\cdot;\Theta_c)$ denotes zero-initialized modules. In addition, the LoRA-based method is considerable. Firstly, concatenate multiple visual condition tokens with noise tokens and text tokens:

$$X^* = [C_T; X; C_V^{(1)}, C_V^{(2)}, ..., C_V^{(n)}], \tag{6}$$

and then input X^* into MMA depicted in Eq. (2), allowing the noise to be restored to the image based on conditions.

Algorithm 1: Framework of DynFusion.

Input:
X: noise image;
C_T: text prompt;
C_V: visual condition images;
t: timestep;
Output:
X_{out}: the predict noise;

$C_V \leftarrow \{C_V^{(1)}, C_V^{(2)}, ..., C_V^{(n)}\}$
// Encode inputs into latent representations separately
$X, C_T, C_V \leftarrow \textbf{Encoder}(X, C_T, C_V)$
for *each i in Blocks[0:-1]* **do**
  // Predict required visual conditions
  $\hat{\mathbf{M}} \leftarrow \mathbf{CAM}(X)$
  if *is_training* **then**
    // Transfer binary mask to attention mask
    $\hat{\mathbf{M}}_{\mathrm{attn}} \leftarrow \mathbf{Transpose}(\hat{\mathbf{M}})$
    $(Q,K,V)_{X,C_T,C_V} \leftarrow \mathbf{Proj}(X, C_T, C_V)$
    // Use masked softmax
    $X, C_T, C_V \leftarrow \mathbf{DMMA}(Q,K,V,\hat{\mathbf{M}}_{\mathrm{attn}})$
  end
  if *is_inference* **then**
    // Remove unselected visual conditions
    $C_V \leftarrow \mathbf{Del}(C_V, \hat{\mathbf{M}})$
    $(Q,K,V)_{X,C_T,C_V} \leftarrow \mathbf{Proj}(X, C_T, C_V)$
    // Use normal softmax
    $X, C_T, C_V \leftarrow \mathbf{DMMA}(Q,K,V)$
  end
  $X, C_T, C_V \leftarrow \mathbf{FFN}(X, C_T, C_V)$
end
$X_{out} \leftarrow \textbf{Decoder}(X)$

A2. Pseudo-code of DynFusion

In this section, we give the pseudo-code algorithm of our DynFusion. The specific training and inference procedure is shown in Algorithm 1.

A3. Implementation Details

Benchmarks. We evaluate the performance of our method on SubjectSpatial200K dataset, which has been partitioned into a training set (139,403 samples) and a testing set (5,827 samples) based on the image quality assessment scores evaluated by ChatGPT-4o [11].

Setup. We use FLUX.1-schnell [3] as our base model and Condition-LoRA weights provided by OminiControl [9, 10] to initialize our Condition-LoRA branches. The rank of trainable Fusion-LoRA module is set to 4. We adopt Adam optimizer with a learning rate of $1e^{-4}$ and weight decay of 0.01. We train our DyFusion for 50,000 steps on 8 NVIDIA A100 GPUs with the resolution of 512×512.

Metrics. To evaluate the generative quality, we compute FID [2] and SSIM [12] between the generated images and the ground truth images. To assess the generative controllability, we calculate MSE for the depth map generation task and F1_score for the canny edge conditional generation. To measure the subject consistency, we compute CLIP-I [5] and DINO [1] between generated images and ground truth images. We use CLIP-T [5] to estimate the text consistency between the generated images and the text descriptions. Additionally, to evaluate the inference overhead, we compare parameters, FLOPs and iteration speed simultaneously. Here, since the iteration speed is easily affected by the deployed equipment and the sampling steps expected by different diffusion models vary, we place particular emphasis on the first two metrics.

A4. More Ablation Studies

A4.1. Results on Subject-Insertion Generation

More qualitative ablation results on Subject-Insertion task are provided here. Fig. A1 explore the effectiveness of the proposed adaptive multi-condition fusion mechanism. Fig. A2 compare the performance of different multi-modal attention mechanisms in our framework. Fig. A3 illustrate the impact of Fusion-LoRA on coordinating different conditions and adapting to dynamic condition strategy. Fig. A4 illustrate the performance and efficiency of the model when condition sparsity is limited.

A4.2. Results on Multi-Spatial Generation

More quantitative and qualitative ablation results on Multi-Spatial task are provided here. Tab. A1 and Fig. A5 explore the effectiveness of the proposed adaptive multi-condition fusion mechanism. Tab. A2 and Fig. A6 compare the performance of different multi-modal attention mechanisms in our framework. Tab. A3 and Fig. A7 illustrate the impact of Fusion-LoRA on coordinating different conditions and adapting to dynamic condition strategy. Tab. A4 and Fig. A8 illustrate the performance and efficiency of the model when condition sparsity is limited.

Figure A1. Qualitative ablation of adaptive multi-condition fusion strategy on Subject-Insertion task.

Figure A2. Qualitative ablation of decoupled multi-modal attention on Subject-Insertion task.

Figure A3. Qualitative ablation of Fusion-LoRA component on Subject-Insertion task.

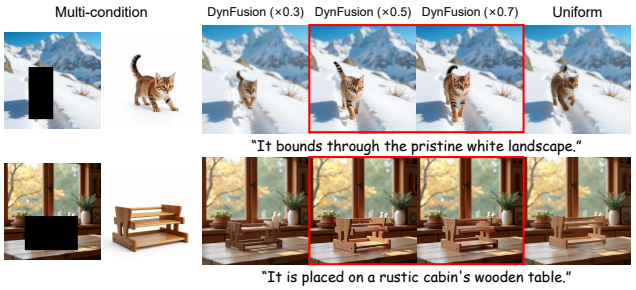

Figure A4. Qualitative ablation of condition sparsity on Subject-Insertion task.

Method	FID ↓	SSIM ↑	F1 ↑	MSE ↓	CLIP-T ↑	FLOPs ↓
Ours.Uniform	6.92	0.63	0.23	182.53	33.21	15.10T
Ours.Sole	6.81	0.63	**0.24**	174.89	33.40	**7.56T**
Ours.Free	**6.52**	**0.66**	**0.24**	**158.97**	**33.44**	8.27T

Table A1. **Quantitative ablation of adaptive multi-condition fusion strategy** on Multi-Spatial task. "Uniform" means all conditions are activated uniformly without any filtering. "Sole" and "Free" indicate selecting the currently most suitable condition and unlimited number of selected conditions, respectively.

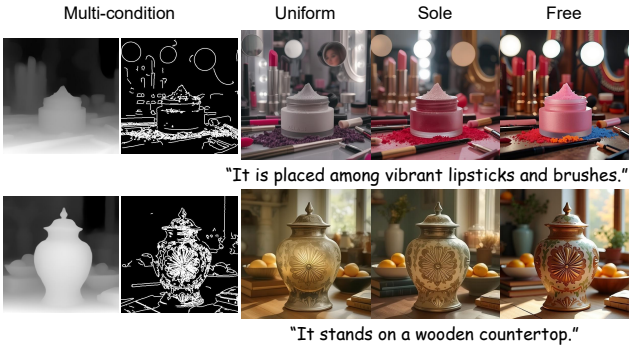

Figure A5. Qualitative ablation of adaptive multi-condition fusion strategy on Multi-Spatial task.

Method	FID ↓	SSIM ↑	F1 ↑	MSE ↓	CLIP-T ↑	AttnOps ↓
Ours w. MMA	6.97	0.61	0.21	225.42	33.11	2.84T / 4.50T
Ours w. CMMA	6.71	0.64	**0.25**	170.53	33.41	2.46T / 3.76T
Ours w. DMMA	**6.52**	**0.66**	0.24	**158.97**	**33.44**	**1.83T / 2.66T**

Table A2. **Quantitative ablation of decoupled multi-modal attention** on Multi-Spatial task. "AttnOps↓" is short for the number of attention operations. Here, the former part is the actual overhead, while the latter is the theoretical overhead under uniform setting.

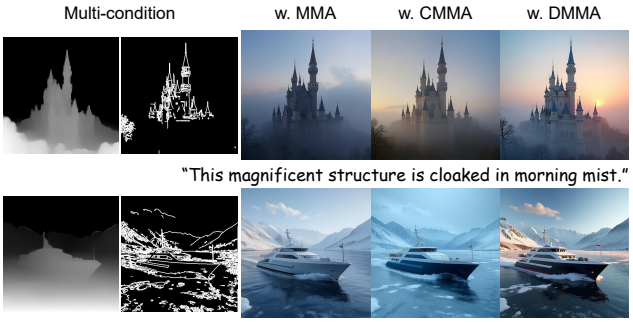

Figure A6. Qualitative ablation of decoupled multi-modal attention on Multi-Spatial task.

Method	FID ↓	SSIM ↑	F1 ↑	MSE ↓	CLIP-T ↑
Ours w/o. Fusion-LoRA	8.74	0.56	0.19	407.02	**33.51**
Ours w. Fusion-LoRA	**6.52**	**0.66**	**0.24**	**158.97**	33.44

Table A3. **Quantitative ablation of Fusion-LoRA component** on Multi-Spatial task. Fusion-LoRA is capable of modulating different quantities and combinations of conditional signals, thereby providing better assistance in generating.

Figure A7. Qualitative ablation of Fusion-LoRA component on Multi-Spatial task.

Method	FID ↓	SSIM ↑	F1 ↑	MSE ↓	CLIP-T ↑	Speed ↑
Ours.Unform	6.92	0.63	0.23	182.53	33.21	2.02it/s
Ours (×0.7)	6.74	0.63	**0.25**	**165.59**	33.28	1.83it/s
Ours (×0.5)	**6.71**	**0.65**	0.23	170.24	**33.48**	2.09it/s
Ours (×0.3)	6.97	0.62	0.21	206.52	33.30	**2.49it/s**

Table A4. **Quantitative ablation of condition sparsity** on Multi-Spatial task. We adjust the sparse loss to obtain models of different sizes.

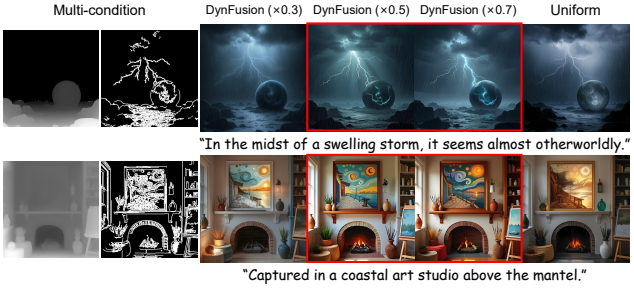

Figure A8. Qualitative ablation of condition sparsity on Multi-Spatial task.

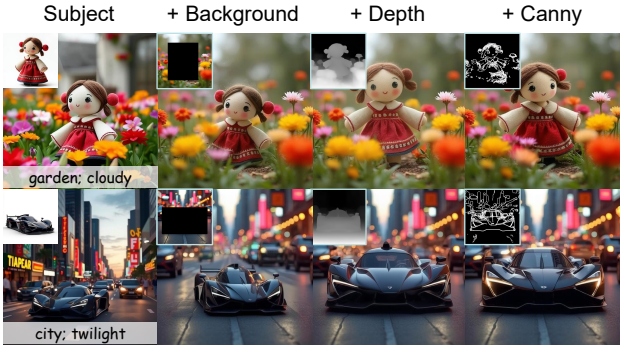

Figure A9. Qualitative ablation of our proposed DynFusion with various multi-condition combinations.

A4.3. Attempt for More Conditions

Our DynFusion imposes no restrictions on the number of visual conditions. As shown in Tab. A5 and Fig. A9, the ad-

Conditions				Generative Quality		Controllability		Subject Consistency		Text Consistency	Inference Overhead		
Subject	Background	Depth	Canny	FID ↓	SSIM ↑	F1 ↑	MSE ↓	CLIP-I ↑	DINO ↑	CLIP-T ↑	Params ↓	FLOPs ↓	Speed ↑
✓		✓		6.21	0.56	-	194.25	94.52	90.70	33.48	**47M**	7.96T	2.04it/s
✓			✓	5.72	0.64	0.25	-	95.33	92.87	33.32	**47M**	8.20T	2.02it/s
✓	✓			4.53	0.80	-	-	97.21	93.14	33.21	**47M**	**7.76T**	**2.09it/s**
✓	✓	✓		4.37	**0.82**	-	146.80	97.27	93.24	33.56	62M	10.65T	1.78it/s
✓	✓	✓	✓	**4.35**	**0.82**	**0.27**	**142.58**	**97.29**	**93.30**	**33.62**	76M	12.98T	1.62it/s

Table A5. Quantitative ablation of our proposed DynFusion with various multi-condition combinations.

dition of more conditions does indeed improve the quality of the generated images to varying degrees. The complementary information among the conditional signals enables more refined control.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[3] Black Forest Labs. Flux: Official inference repository for flux.1 models, 2024. Accessed: 2024-11-12. 2

[4] Yanjie Pan, Qingdong He, Zhengkai Jiang, Pengcheng Xu, Chaoyi Wang, Jinlong Peng, Haoxuan Wang, Yun Cao, Zhenye Gan, Mingmin Chi, et al. Pixelponder: Dynamic patch adaptation for enhanced multi-conditional text-to-image generation. *arXiv preprint arXiv:2503.06684*, 2025. 1

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[7] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[8] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 1

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

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

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

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2