Intermediate Adapter: Efficient Alignment of Text in Diffusion Models

Anonymous ACL submission

Abstract

Diffusion models have been widely used for text-to-image generation tasks. However, state-of-the-art models still fail to align the generated visual concepts with high-level semantics in a language such as object count, spatial relationship, etc. We approach this problem from an architectural perspective and investigate how conditioning architecture can affect vision-language alignment in diffusion models. We propose a new conditioning architecture named Intermediate Adapter to improve text-to-image alignment, generation quality, as well as training and inference speed for diffusion models. We perform experiments on the text-to-image generation task on the MS-COCO dataset. We apply Intermediate Adapters on two common conditioning methods on a U-ViT backbone. For both end-to-end training and fine-tuning of pretrained diffusion models, Our method boosts the CLIP Score, FID, and human evaluation results of the generated images, with 20% reduced FLOPs, and 25% increased training and inference speed.

1 Introduction

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Dhariwal and Nichol, 2021; Rombach et al., 2022a) have emerged as a dominant framework for generating images from natural language. By leveraging prealigned text embeddings such as CLIP, diffusion models can generate high-definition images from text prompts (Ramesh et al., 2021, 2022; Betker et al., 2023; Rombach et al., 2022b; Podell et al., 2023). Most text-to-image diffusion models use concatenation or cross-attention to merge the pretrained CLIP text embedding into the image-only diffusion model. This approach's core issue is the inherent gap between the CLIP and diffusion objectives. The CLIP aligns the text embeddings and image features, but the diffusion training takes in different levels of noisy image features. Although this misalignment can be reduced by fine-tuning the CLIP embedding in the diffusion training process, this approach complicates the already complex diffusion training. In addition, most existing diffusion models follow a simple design of adding text conditions on all levels of the backbone architecture. This design potentially introduces redundant text guidance with additional computing complexity. Although some works (Zhao et al., 2023) have experimentally discovered that it does not harm the performance to trim certain attention layers for efficiency, the reason behind it has not been thoroughly studied. Thus in our work, we carefully examine the text-image interactions in diffusion backbones, and based on this, design a special mechanism to align the text embedding to the image diffusion task efficiently.

In this paper, We investigate a specific type of ViT-based diffusion backbone. By examining the text-to-image and image-to-image attention maps at different layers, we discover that semantic information from text prompts provides more guidance near bottleneck layers whereas fusing text information at earlier or later layers provides minimum guidance. Based on this observation, We propose a new conditioning architecture named Intermediate Adapter. This method has two major design components: 1. removes the text-conditioning mechanism from the early and late layers and 2. adds additional text-only transformer adapter layers that are trainable in the diffusion process. Analytical experiments indicate that component 1 improves the efficiency of the text-image cross-attention mechanism, and reduces the interference between image and text, leading to higher quality generation. Component 2 improves the text-image alignments of the generated images. When combined, we see a large margin of improvement in all evaluation metrics.

As a result, our proposed Intermediate Adapter can enhance a diffusion model to generate better quality and more text-aligned images, especially for high-level semantics such as accurate object

Figure 1: Text-to-image generation on MS-COCO validation captions. U-ViT (top) vs. U-ViT + Intermediate Adapter (bottom). Images are generated using classifier-free guidance based on text prompts. Each text prompt contains a type of high-level semantics in each column. Our method enhances U-ViT to generate better quality samples that align better with the text prompt and are more efficient in training and inference.

count, compound concepts, relationships between multiple objects, rare concepts, and entangled concepts, etc. (see Figure 1 for a few generated samples). Our method also makes diffusion models more efficient, requiring less computing, memory, training time, and inference time.

2 Related Work

ViT-based diffusion model backbones have been explored recently (Peebles and Xie, 2023; Bao et al., 2023a). They bring several large-scale applications in the text-guided generation domain (Karaarslan and Aydın, 2024; Rombach et al., 2022b) and multimodal generation domain (Bao et al., 2023b). These models leverage different mechanisms to fuse the text guidance to the diffusion model, with simple concatenation and cross-attention being the two commonly used mechanisms. As for text-image alignment, most existing works try to improve the alignment from a training perspective including finetuning with augmented data (Paiss et al., 2023; Betker et al., 2023), introducing additional alignment guidance (Wu et al., 2023), etc. Quite differently, we approach this problem from an architectural perspective to enable better alignment without additional data. Regarding efficient diffusion models, recent works include reducing sampling steps using an efficient sampler (Song et al., 2020a; Lu et al., 2022a,b), consistency training and distillation (Song et al., 2023; Luo et al., 2023), or reusing calculations across timesteps (Zhang et al., 2024). These works focus more on algorithmic efficiency, while our work focuses on reducing architectural redundancy. Adapters are commonly used in diffusion models to provide additional control in generation (Zhang et al., 2023; Mou et al., 2023; Ye et al., 2023). However, in this study, we explore an extra functionality of adapters to reduce the inference between text conditions and generated images, leading to improved alignment between them.

3 Preliminaries

3.1 Diffusion Models

Diffusion models (Song et al., 2020b; Ho et al., 2020) are generative models that learn to generate new samples from noise by approximating the score function of a data distribution $p(x)$ using a neural network. The score function is defined as the gradient of the log-probability density of the training data points:

$$S(x) = \nabla_x \log p(x). \tag{1}$$

Training. Given a dataset with data distribution $p(x)$, the training involves two steps:
Noise addition. Given a data sample x, we progressively inject Gaussian noise over T steps until reaching a full noise x_T. This process can be formalized as follows:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t, \tag{2}$$

where $\epsilon_t \sim \mathcal{N}(0, I)$, $t = 1, \ldots, T$, $x_0 = x$, and

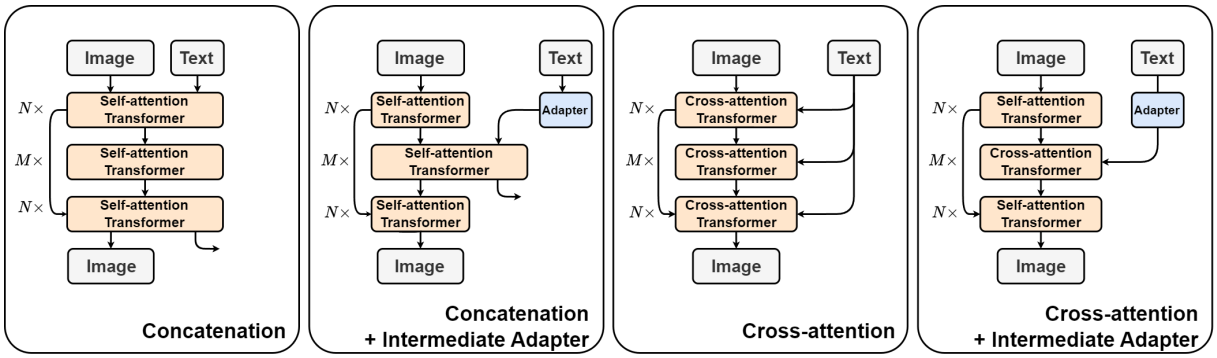

Figure 2: Two conditioning methods using a U-ViT backbone, with and without Intermediate Adapter. We only show 3 groups of transformer blocks for simplicity. In our experiments, all setups have 13 layers of transformer blocks, with $N = 4$ and $M = 5$. Time embeddings, pre-processing layers, and post-processing layers are omitted for simplicity. In practice, the time embedding is concatenated with the image input and follows the same path.

$\{\beta_t\}_{t=0}^{T}$ is a noise schedule.

Score function learning. The training objective is to minimize the discrepancy between the true score function $S(\cdot)$ and its neural network approximation $s_\theta(\cdot)$:

$$L(\theta) = \mathbb{E}[(S(x_t) - s_\theta(x_t))^2]. \quad (3)$$

Sampling. The sampling process is in a reverse direction. We first initialize the noisy sample x_T from a standard Gaussian distribution. We then apply the learned score function to denoise x_T over $t \in [T, 1]$ steps to gradually remove noise and generate a sample:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\beta_t}} s_\theta(x_t)) + \sqrt{\beta_t}\epsilon \quad (4)$$

Classifier-free guidance (CFG). We can also model the conditional score function $S(x_t|y)$ by approximating the unconditional score function $\nabla_{x_t} \log p(x_t)$ and the joint score function $\nabla_{x_t} \log p(x_t, y)$ simultaneously to enable conditional generation:

$$S(x_t|y) = (1+\omega)\nabla_{x_t} \log p(x_t, y) - \omega \nabla_{x_t} \log p(x_t), \quad (5)$$

where ω is the CFG scale that controls the strength of guidance. The conditional generation in our study is achieved through classifier-free guidance from caption y to image x, while unconditional generation uses the same approach with an empty caption embedding.

3.2 Conditioning Methods

Our major contribution is to enhance the current conditioning approaches with the Intermediate Adapter. In our empirical exploration, we mainly focus on two common conditioning methods used by SOTA text-to-image diffusion models:

Concatenation. Text, image, and timestamps are all processed as tokens and concatenated. They are fed into a self-attention transformer as a long sequence. (See Figure 2 sub-figure 1.)

Cross-attention. The image self-attention is joined by the cross-attention from the conditioning text. (See Figure 2 sub-figure 3.)

We only use these two common approaches to show that our method is applicable and effective on different conditioning approaches. Although AdaNorm has also demonstrated effectiveness in recent diffusion models (Peebles and Xie, 2023; Crowson et al., 2024), its original version does not support long conditioning texts, which restricts its use in our text-to-image generation task.

4 Methodology

We first introduce the base diffusion model and the backbone we use in sections 4.1 and 4.2. Then we introduce our Intermediate Adaptor in sections 4.3 and 4.4, each focusing on one of the two components: intermediate fusion and text adapter.

4.1 Latent Diffusion

Latent Diffusion Models (LDMs) (Rombach et al., 2022b) operate directly in the latent space of pre-trained image features. We use a stable diffusion KL-based autoencoder to encode an input image into the latent space and decode the denoised latent space representation back to the input image space. For text embeddings, we use the CLIP embedding with ViT-L-14. These models are frozen during diffusion model training.

4.2 Diffusion Backbone Model

Evidence suggests that under a diffusion model setting, segmentation networks with long skip connections are essential to the efficient learning of discrete time ODE (Huang et al., 2024). When long skip connections are used, distant network blocks can be connected to aggregate long-distant information and alleviate vanishing gradient. For this reason, we choose U-ViT-Small (Bao et al., 2023a) as our baseline backbone. The model proposed in the orginal paper uses concatentation to merge the text information. (Figure 2 sub-figure 1). On top of this, we also study another cross-attention setting (Figure 2 sub-figure 3). These settings are constructed by only changing the architecture without modifying the training and inference setups.

4.3 Intermediate Fusion

In multimodal fusion, intermediate fusion refers to a fusion that occurs at an intermediate level. This way different modality data are allowed to preprocess in a single modality manner before joining a shared latent space. We borrow the same idea in the context of diffusion models. In our method, we remove the text conditioning at the beginning and the end of the diffusion model as shown in Figure 2 sub-figures 2 and 4. In the specific setup, we remove 4 layers of text-conditioning mechanism each from the beginning and end of the 13 layers of diffusion backbone, keeping only 5 intermediate layers text-guided. This choice is made from the observations in Section 5.6 that the text guidance is weak at the early and late stages of the backbone. Removing these text conditionings reduces the text-related attention calculations, thus improving the speed of the model.

4.4 Text Adapter

Instead of directly introducing pretrained CLIP text embedding in the diffusion backbone, we add a preprocessing adaptor transformer layer for the text embedding as shown in Figure 2 sub-figures 2 and 4. This layer allows the text embedding to be fine-tuned to better align with the diffusion task. The adapter is a one-layer multi-head self-attention transformer (Figure 3).

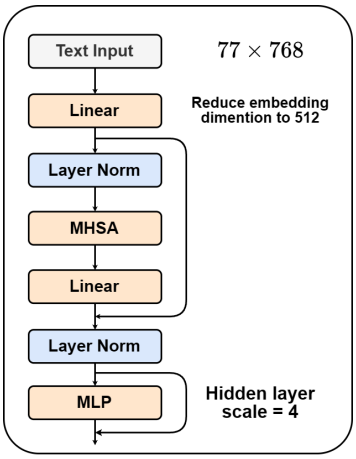

Figure 3: The text adapter architecture. Here Multi-head Self-attention (MHSA) and Multi-layer Perceptron (MLP) follow the default implementations used in ViT.

5 Experiments

5.1 Dataset and Training Settings

In our experiments, we use the MS-COCO (Lin et al., 2015) train and validation datasets to train and evaluate the performance of our model. For our training configuration, we train all models for 1 million steps and use a batch size of 256. We use the AdamW optimizer, with a learning rate of 0.0002, weight decay of 0.03, and beta parameters set to (0.9, 0.9). We incorporate a warm-up phase of 5000 steps to adjust the learning rate. The ViT model takes image features with a channel of 4, both spatial dimensions of 32 and an image patch size of 2. All attention mechanisms use an embedding dimension of 512 and 8 attention heads. CLIP embedding has 77 tokens each with a dimension of 768, and is transformed to a dimension of 512 using a linear layer to align with the transformer input. For classifier-free guidance, we use a probability of 0.1 for unconditional training.

5.2 Evaluation Metrics

Quantitative evaluation. We use FID, and CLIP Score as our quantitative metrics. To generate the score we select 30000 captions from the MS-COCO validation set and the corresponding generated images from our text-to-image models. For CLIP Score we use the CLIP version CLIP-ViT-L-14.

Human evaluation - object count. We choose a challenging generation aspect even for most of the foundation text-to-image models - matching object count, where we require the model to generate the same amount of objects as described in the prompts. We select four objects - bus(es), sheep, person(people), and apple(s). These four are selected since they represent 4 different plural forms and 4 categories (human, animal, fruit, human-

Table 1: Comparative results on text-to-image generation and alignment metrics. The baseline U-ViT corresponds to the first setting. For training speed, an iteration (iters) is a full forward-backward pass on an RTX-4090 GPU with a mini-batch size 256. GFLOPs are calculated on a single forward pass of the model at a timestamp.

Conditioning Method	Model	FID-30K ↓	CLIP Score ↑	Training iters/s	GFLOPs
Concatenation	U-ViT	5.98	0.584	1.81	29.56
	U-ViT + IA	**5.77**	**0.588**	**2.31**	**25.84**
Cross-attention	U-ViT	6.48	0.575	2.54	23.82
	U-ViT + IA	**5.68**	**0.588**	**2.74**	**23.66**

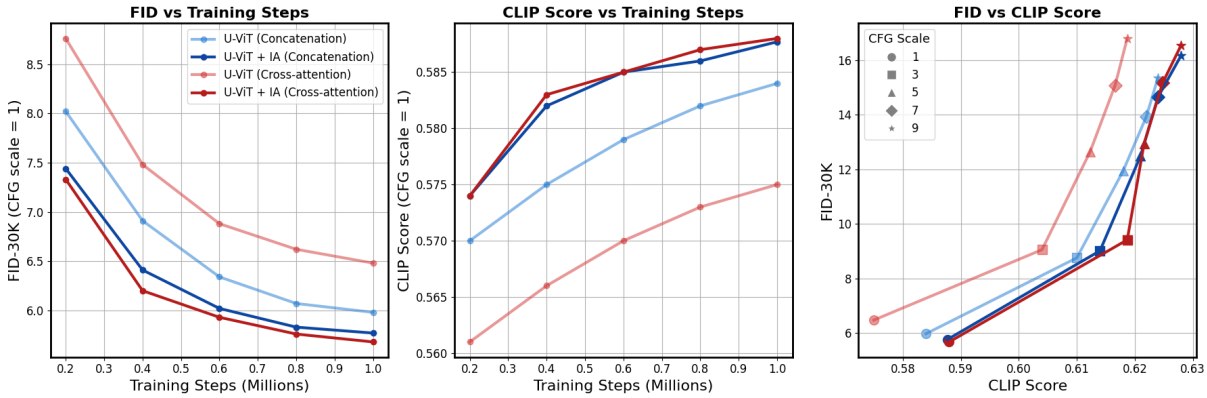

Figure 4: Evaluation during training and FID-30K vs CLIP Score at different CFG scales. settings with IA show improved generation quality and text-image alignment compared to their early fusion counterparts. CLIP Score is measured on 30K pairs using CLIP-ViT-L-14.

made object). We use 5 words of count - a(an), two, three, four, five. Since larger numbers are rare in MS-COCO training captions, we restrict our study to small numbers. We generate 10 images for each object-count pair (20 pairs) and let evaluators count the number of target objects in the generated image. Then we use the average error (AE) and average match ratio (AMR) to evaluate the performance, based on the evaluators' counts C_{eval} and the prompts' counts C_{prompt}:

$$\text{AE} = \frac{\sum_{i=1}^{n} |C_{i,\text{human}} - C_{i,\text{prompt}}|}{n} \quad (6)$$

$$\text{AMR} = \frac{\sum_{i=1}^{n} \mathbb{I}(C_{i,\text{human}} = C_{i,\text{prompt}})}{n} \quad (7)$$

Human evaluation - preference score. In addition, we ask 5 evaluators to provide a preference ranking from 1 to 4 on the overall quality of images generated by each model given captions from the evaluation set. We use the same random seed and prompts for all models and provide the generated images with prompts to the evaluators. The human evaluators are told to skip any group of samples if the ranks are hard to call. 100 captions are evaluated by 5 evaluators, with a maximum of 500 scores for each setting. We assign 4, 3, 2, and 1 scores to rank 1, 2, 3, and 4 respectively, and calculate the average score for each model setting.

5.3 Results

We selected U-ViT (Bao et al., 2023a) as our baseline model since it has the best MS-COCO FID score among dedicated diffusion models with a manageable size. We first compare the performance between 2 different conditioning methods with their counterparts with intermediate adapters in Table 1. We observed that the intermediate adapter (IA) improves the FID, CLIP Score, training speed, and FLOPs of the base models.

Next, we visualize the FID (Figure 4, left), CLIP Score (Figure 4, middle) during training, and FID vs. CLIP Score at different CFG scales (Figure 4, right). We find that the models with intermediate adapters exhibit better FID and CLIP scores throughout the training. They also show a better trade-off between CLIP Score and FID. Among all four settings, a cross-attention U-ViT with an intermediate adapter has the best FID, CLIP score, lowest GFLOPs, and fastest training.

We then select 12 random captions and generate

Figure 5: Generated samples comparison between the baseline U-ViT(top) and U-ViT with an intermediate adapter across 12 validation prompts. (Best viewed when zoomed-in.)

images with a CFG scale of 3, which is an elbow point in the FID vs. CLIP Score curve. We show the baseline (top) compared with the one with an intermediate adapter (bottom) in Figure 5. We observe that the generated images are more spatially consistent and more aligned with the text prompts.

We then show that our intermediate adapter boosts U-ViT against several foundation models and dedicated models (Table 2). Our model can reach the best text-image alignment performance and comparable image quality to all models with a relatively small model size.

Table 2: Performance of text-to-image diffusion models.

Model	FID-30K ↓	CLIP Score ↑
Foundation Models Zero-shot on MS-COCO		
Imagen	7.27	∼0.29
Stable Diffusion	8.59	0.325
Models Trained/Finetuned on MS-COCO		
VQ-Diffusion	19.75	-
Frido	8.97	-
U-ViT	5.98	0.584
U-ViT+IA(ours)	**5.68**	**0.588**

5.4 Human Evaluation

Object count. The results are shown in Figure 6. In the left four figures, for 18 out of 20 object-count pairs, U-ViT using concatenation with intermediate adapters generates objects with more human-aligned count compared to the baseline. For 14 out of 20 object-count pairs, U-ViT using cross-attention with intermediate adapters generates objects with more or equal human-aligned count compared to the baseline. In the top right figure, the average error of models with intermediate adapters is consistently lower than the baselines. In the bottom right figure, the average match ratio of models with intermediate adapters is consistently higher or on par with the baselines. All of the above results show that the intermediate adapter improves the count alignment in the generation regardless of the conditioning method.

Preference score. The results are shown in Figure 7. 287 scores are collected after removing invalid scores and those are too difficult to call by the human evaluators. The score is consistent with our FID and CLIP Score evaluation, with cross-attention U-ViT with an intermediate adapter achieving the highest score, and concatenation with an intermediate adapter coming second. All settings with intermediate adapters outperform their corresponding baseline models.

5.5 Ablations

We also apply the two components of the intermediate adapter separately and study their individual contribution to the FID, and CLIP Score. In Figure 8, we show that fusing the text embedding only in the middle of the diffusion backbone is the major source of FID improvement. This is expected since this setting has image-only skip connections that can maintain spatial consistency at the upsampling layers. But this will negatively impact the CLIP Score. Adding a text adapter learns more aligned text embeddings, which is the major source of improved CLIP Score, but this will impact the model efficiency in terms of increased FLOPs and reduced training speed. When the two methods are combined, we achieve improvements in all metrics. We see that these two components compensate for each other's weaknesses while maintaining their respective advantages in the intermediate adapter.

5.6 Analysis

Layer-wise Attention Maps. To better visualize the text-image alignment across the model layers, we analyze the average attention maps of all timesteps during the diffusion process. In Figure 9 we show the comparison of U-ViT and U-ViT with an intermediate adapter. The text-to-image attention maps in both early and late layers indicate

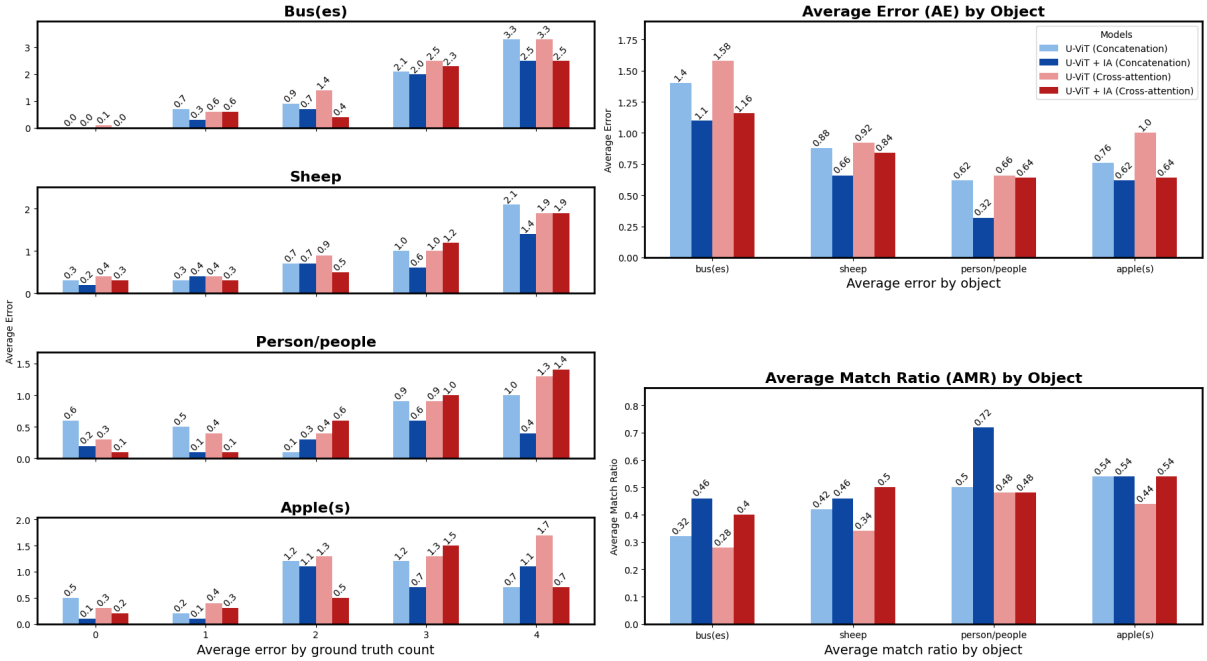

Figure 6: Human evaluation on object count. Lighter colors represent the baseline U-ViTs, while darker colors represent corresponding models with an intermediate adapter. The left four figures are the average error given different ground truth counts, where the x-axis is the ground truth. Each figure corresponds to an object. The right top figure is the average error across all counts for different objects. The bottom right figure is the average match ratio for each object. The plots indicated lower average count errors and higher matching counts of intermediate fusion.

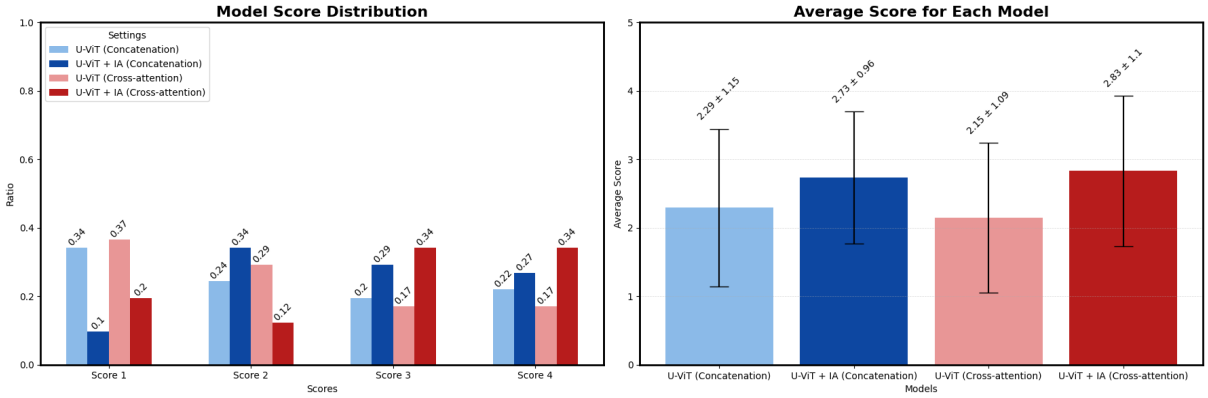

Figure 7: Human evaluation on general quality of generation. Models with intermediate adapters (deep blue and red) have more frequent high scores (left). They also have higher average scores than the corresponding baselines (right).

a more uniformly distributed pattern than intermediate layers, suggesting that text guidance is less focused and effective in early and late layers. Besides this observation, we observe that the early and late layers attend more to the border of the latent image due to the padding added in the convolutional layers in the autoencoder model. To reduce the influence of such padded borders, we removed the border so that the later rank analysis could reflect more semantic guidance.

Rank Analysis on Adjusted Attention Map To quantify the influence of text guidance on image features, we conducted SVD on the text-to-image attention map matrices and analyzed their rank property in Figure 9 bar charts below the attention maps. Since the softmax function is applied to the attention map QK^T, each layer is normalized thus providing a fair comparison across layers. We see that U-ViT models have relatively low-rank text-to-image attention maps with smaller singular values at all layers, especially the layers away from the middle. On the other hand, the one with an interme-

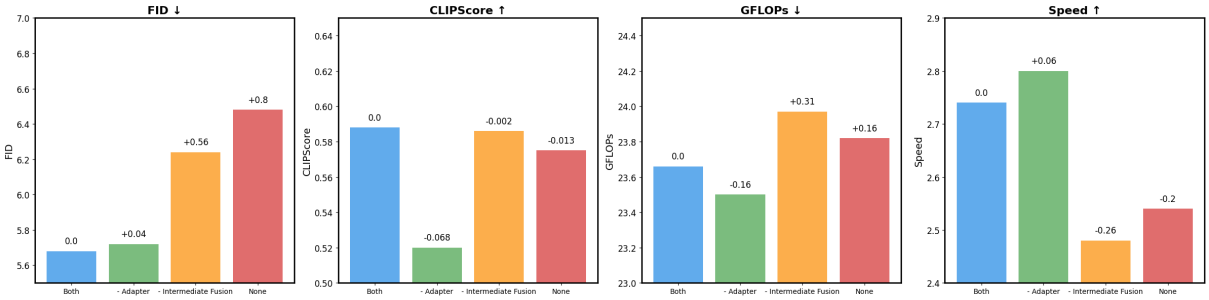

Figure 8: Ablation Study. We show the effect of the adapter and intermediate fusion separately by removing the corresponding components (green and orange), compared to the joint effect (red) and baseline (red). Despite being more efficient by removing the adapter (green), it impacts the CLIP-Score and FID negatively. When both components are present(blue), the model shows a better balance between performance (sub-figure 1, 2) and efficiency (sub-figure 3, 4).

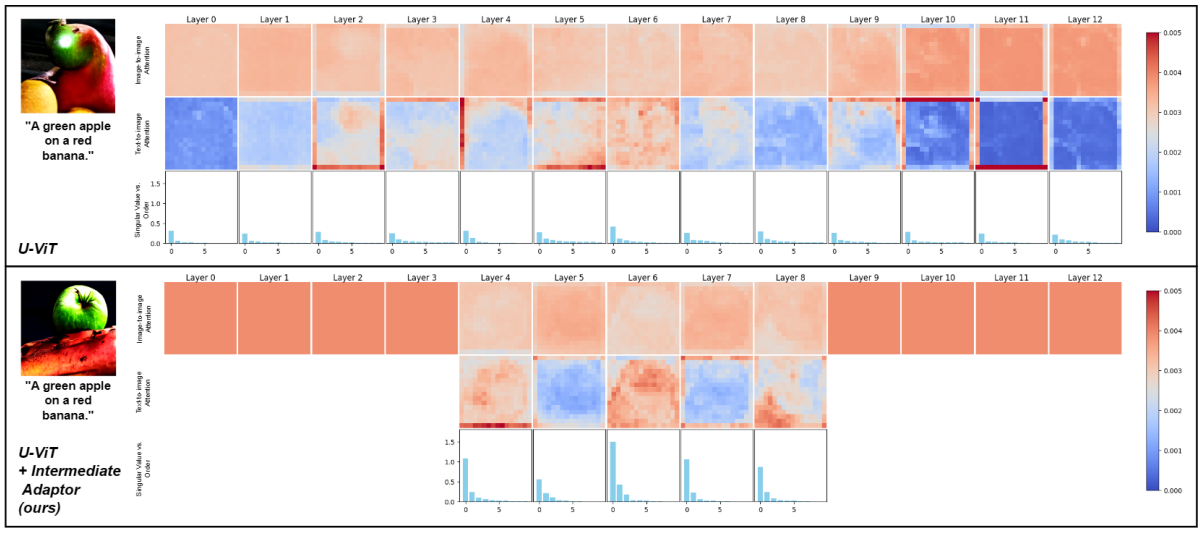

Figure 9: Attention maps and singular value analysis. For each setting, the first row is the image-to-image attention, the second row is the text-to-image attention, and the third row is the singular values of the first 10 orders from the text-to-image attention maps. An intermediate adapter removes the low-information text-to-image attention at the early and late levels. This reduces the interference between image and text at these levels and improves the information capacity of the text-to-image attention at the middle layers.

diate adapter has high singular values. The analysis indicates that low-capacity text-to-image attention occurs in the early and end layers of a diffusion backbone, whereas most of the text information is processed around the bottleneck. This justifies our presumption of redundant text guidance in U-ViT. Additionally, the comparison of singular values around the intermediate layers proves that the elimination of the early and end fused layers never hurts the effectiveness of guidance. It instead boosted the guidance in the intermediate layers. Thus, we can potentially improve model efficiency without damaging the semantic control of text. This observation aligns with the experiment results.

6 Conclusion

In this study, we presented an effective architecture for enhancing text-to-image diffusion models by leveraging an intermediate adapter mechanism for text conditioning. Our experiments and analyses on the MS-COCO dataset demonstrate that this method outperforms traditional architectural design in aligning visual concepts with language, improving generation quality, and enhancing the efficiency of the training and inference. More generally, our findings suggest that the placement and preprocessing of text embeddings within diffusion models play a critical role in the performance and efficiency of text-to-image generation tasks. This provides a direction for large foundation models to a more efficient and text-aligned design.

7 Limitations

Other conditioning strategies. The main focus of this paper is to investigate the influence of an intermediate adapter on a diffusion backbone, in terms of text-image alignment, generation quality, and computational efficiency. Admittedly, some other conditioning methods such as the AdaNorm used by DiT are not explored in this paper. We reasonably argue that the intermediate fusion can be transferred with ease to other unexplored conditioning strategies since the approach resolved the issue of less efficient text guidance caused by joining image and text at early and late levels of a diffusion model.

Pretrained model fine-tuning. Since our method uses an adapter, it can be fine-tuned on pretrained foundation models to replace text conditioning. From our experiments, full fine-tuning of a pretrained U-ViT can achieve comparable performance with the end-to-end training with only 5% of total steps. However, due to the limited computation resources, its application to foundation models is not discussed in this paper. The issue is that our method aims to learn less interfered features of the image and trainable embeddings for language, which require changes in all layers of the diffusion backbone. This requires the full fine-tuning of a pretrained model. However, full fine-tuning of a foundation model is beyond the scope of this work. The main focus of this work is to show that multimodal information fusion in diffusion models should follow an intermediate fusion design, where the conditions should be preprocessed jointly within the diffusion process. The goal of this work is to inspire the next-generation foundation model design to follow a similar design for better generation quality and condition-following.

References

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. 2023a. All are worth words: A vit backbone for diffusion models. *Preprint*, arXiv:2209.12152.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. 2023b. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv preprint arXiv:2303.06555*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z. Kaplan, and Enrico Shippole. 2024. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. *ArXiv*, abs/2401.11605.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion models beat gans on image synthesis. *Preprint*, arXiv:2105.05233.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Preprint*, arXiv:2006.11239.

Zhongzhan Huang, Pan Zhou, Shuicheng Yan, and Liang Lin. 2024. Scalelong: Towards more stable training of diffusion model via scaling network long skip connection. *Advances in Neural Information Processing Systems*, 36.

Enis Karaarslan and Ömer Aydın. 2024. Generate impressive videos with text instructions: A review of openai sora, stable diffusion, lumiere and comparable models. *Authorea Preprints*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022a. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022b. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *ArXiv*, abs/2211.01095.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *Preprint*, arXiv:2310.04378.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. *Preprint*, arXiv:2302.12066.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. *Preprint*, arXiv:1503.03585.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020a. Denoising diffusion implicit models. *arXiv:2010.02502*.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models. In *International Conference on Machine Learning*.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020b. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score: Better aligning text-to-image models with human preference. *Preprint*, arXiv:2303.14420.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Wentian Zhang, Haozhe Liu, Jinheng Xie, Francesco Faccio, Mike Zheng Shou, and Jürgen Schmidhuber. 2024. Cross-attention makes inference cumbersome in text-to-image diffusion models. *arXiv preprint arXiv:2404.02747*.

Yang Zhao, Yanwu Xu, Zhisheng Xiao, and Tingbo Hou. 2023. Mobilediffusion: Subsecond text-to-image generation on mobile devices. *Preprint*, arXiv:2311.16567.