MaxFusion: Plug&Play Multi-Modal Generation in Text-to-Image Diffusion Models

Nithin Gopalakrishnan Nair[1], Jeya Maria Jose Valanarasu[2], and Vishal M Patel[1]

Johns Hopkins University[1], Stanford University[2]
{ngopala2, vpatel36}@jhu.edu, jmjose@stanford.edu
https://nithin-gk.github.io/maxfusion.github.io/

Abstract. Large diffusion-based Text-to-Image (T2I) models have shown impressive generative powers for text-to-image generation and spatially conditioned image generation. We can train the model end-to-end with paired data for most applications to obtain photorealistic generation quality. However, to add a task, one often needs to retrain the model from scratch using paired data across all modalities to retain good generation performance. This paper tackles this issue and proposes a novel strategy to scale a generative model across new tasks with minimal computation. During our experiments, we discovered that the variance maps of intermediate feature maps of diffusion models capture the conditioning intensity. Utilizing this prior information, we propose MaxFusion, an efficient strategy to scale up text-to-image generation models to accommodate new modality conditions. Specifically, we combine aligned features of multiple models, bringing a compositional effect. Our fusion strategy can be integrated into off-the-shelf models to enhance their generative prowess.

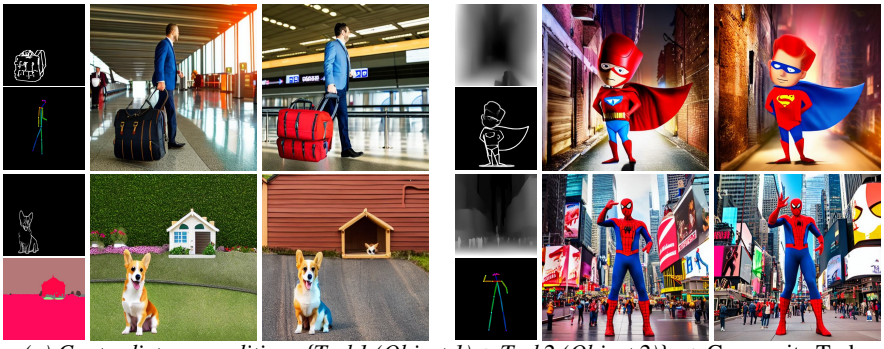
Keywords: Multimodal conditioning · Plug and Play · Diffusion Models

1 Introduction

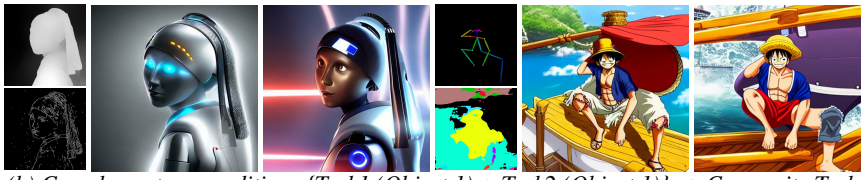
Denoising Diffusion Probabilistic Models (DDPMs) [9] have revolutionized the content creation industry by delivering photorealistic quality in image [2, 8, 34, 36], video [7, 10], and 3D generation tasks [22, 31]. These models translate textual descriptions or class labels into precise visual representations, thanks to the training that conditions them on text or labels [28, 29]. The remarkable capabilities of recent text-to-image (T2I) diffusion models [2, 34] have expanded further with the integration of additional control signals, such as instructions [3] or style-based conditioning [23, 35, 50]. The integration of additional control allows for creating images that more closely align with user preferences by enabling more nuanced control through multi-tasking capabilities, enabling multiple conditions to be given to the model simultaneously.

However, training a generative model to accommodate multiple conditions presents significant challenges, primarily due to the need for paired multi-modal data [20, 47]. Moreover, achieving satisfactory performance from multi-modal training often necessitates prolonged training periods. Transfer learning has emerged as a popular solution for multi-modal generation of T2I Diffusion models [42, 44], allowing large text-to-image models trained on extensive datasets like LAION-2B [37] to undergo fine-tuning for downstream tasks, including image semantics [47, 48] and instruction-based

synthesis [3], in a comparatively shorter time using smaller datasets (around 10M images) [32]. Consequently, with sufficient multi-modal paired data and robust computing resources, fine-tuning a pre-trained text-to-image model to incorporate a broad range of conditions becomes feasible [17, 35]. However, this approach hinges on a fixed set of initial conditions and a strategic training data selection. A notable drawback is that fine-tuning with smaller datasets can lead to catastrophic forgetting [15, 47], where the model loses previously acquired knowledge from the original T2I task.



(a) Contradictory conditions {Task1 (Object 1) + Task2 (Object 2)} → Composite Task



(b) Complementary conditions {Task1 (Object 1) + Task2 (Object 1)} → Composite Task

Fig. 1: Applications of MaxFusion: **(a)** Contradictory condition composition: We show cases where the generated images can easily bring in effects from multiple conditions that otherwise need extremely descriptive conditioning. **(b)** Complementary condition composition: Multimodal generation combining details from different modalities but from the same image.

Recent works such as ControlNet [49] and T2I Adapter [26] tackled this problem by introducing task-specific trainable parameters to include additional spatial conditioning for diffusion models. Both these approaches showed impressive image synthesis quality with good coherence to the structural conditioning. However, adapting ControlNet and T2I Adapters to multiple tasks requires either training from scratch or manual parameter tuning across the intermediate outputs of multiple adapters, which accounts for a non-trivial increase in compute time and memory. Hence, synthesizing images/videos satisfying multiple conditions remains a challenge [4]. Progressing towards a solution for multi-conditioning, a recent work, Unicontrol [32], trained a generative model with paired data across diverse conditional modalities [11]. Through the multitasking-based training paradigm, Unicontrol achieved better performance for individual spatial conditioning tasks than Controlnet. Unicontrol has been trained with large-scale curated data [37] across various modalities to achieve good visual synthesis quality. However, the challenge associated with such an approach is that in order to add a new conditioning task, the models must undergo retraining, incorporating data not only from the new

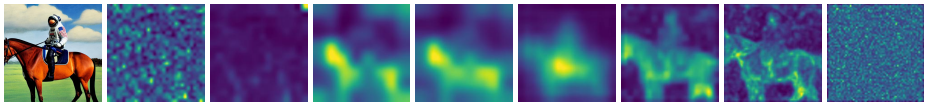


Fig. 2: Figure illustrating variance maps of intermediate features of encoder and decoder for the text prompt "An astronaut riding a horse" for the 5^{th} timestep of diffusion.

task but also from all preceding tasks [27, 32]. We refer to this as the *"scaling issue"* signifying a rise in training complexity when integrating an extra task into a pre-trained model. A training-free solution for multi-conditioning in the literature is compositional generation [5, 6, 21]. However, compositional algorithms impose a linear overhead on inference memory. To this end, we address the question.

"Can we enable a single diffusion model to scale efficiently across multiple diverse tasks without retraining?".

In this paper, we propose a training-free solution for multi-modal generation. We achieve this by utilizing a novel feature fusion scheme that leverages distinct single-task models for compositional generation. Our feature fusion criteria are motivated by an observation regarding the intensity of the conditioning features and the variance maps of intermediate layers of a text-to-image diffusion model. We present an example in Figure 2. Based on this observation, we derive a criterion from feature variance maps across different layers of a diffusion model. We propose MaxFusion - a simple and efficient feature fusion algorithm that allows scaling up T2I models to multiple tasks simultaneously, enabling zero-shot multi-modal generation powers to T2I models. Figure 1 presents different applications of our method. In Figure 1(a), the top row illustrates the case of zero-shot multi-modal generation where two models trained for individual tasks are combined only during inference time to obtain multi-modal generation. These cases illustrate cases where we can combine different modalities of different objects to create a compositional scene. For example, in the first case, we add the pose of a person to a sketch of an object and create a scenario where it can interact. Other examples also show similar multi-modal generations with contradictory condition composition of pose, sketch, depth maps and semantic maps. Our method can also be utilized for multi-modal conditioning of the same object to add complementary information, and this can be seen in Figure 1(b). For example, we can see that $Depth, Edges$ conditions the same image with complementary information. Our proposed merging mechanism is versatile and can be applied atop any off-the-shelf diffusion models like ControlNet and T2I-Adapter, enabling compositional conditioning tasks and extending their generative capabilities. We perform extensive experiments to show the effectiveness of the proposed method. We present results on: (i) Multi-modal generation using multiple conditions describing different levels of semantics on the same spatial locations in an image. (ii) Zero-shot generation, where we utilize individual models trained for different tasks and show the combined task without retraining.

In summary, the following are the main contributions of this work:

- We tackle the need for training with paired data for multi-task conditioning using diffusion models.
- We propose a novel variance-based feature merging strategy for diffusion models.

- Our method allows us to use the combined information to influence the output, unlike individual models limited to a single condition.
- Unlike previous solutions, our approach is easily scalable and can be added to off-the-shelf models.

2 Related Works

2.1 Scaling Diffusion Models to Additional Conditions

Denoising Diffusion Probabilistic Models (DDPMs) [9] have achieved remarkable results in a variety of tasks, including text-conditioned image generation [28, 34, 36], video generation [8, 10], and multi-modal generation [34, 35]. Methods like ControlNet [49] and T2I Adapter [26] have been developed to extend the capabilities of these large, pre-trained diffusion models, enabling them to accommodate additional input conditions. These conditions influence the spatial semantics of the generated images by using paired $image, text, conditioning$ triplets, synthetically created for this purpose. Various semantic details, including edge maps, poses, depth, and segmentation maps, serve as conditioning inputs for the model.

To add a new task, a task-specific encoder carrying the original encoder's weights is seamlessly integrated into the network. This encoder then connects to convolutional layers initialized at zero, aiming to maintain consistent intermediate latent outputs early in training, thus ensuring a stable foundation for effective learning. Only the parameters of this additional encoder undergo training to introduce the new condition, avoiding the retraining of pre-existing model weights. This approach is crucial as retraining pre-trained weights could lead to catastrophic forgetting [15], a significant risk that ControlNet and T2I-Adapter skillfully navigate to preserve the original model's visual synthesis quality without any performance degradation.

2.2 Model Merging

Recent advancements in Large Language Model (LLM) research have introduced the innovative concept of model merging [6, 24, 40, 45, 46], a process where weights from multiple pre-trained models for the same task are combined. This technique aims to create a superior model that outperforms the individual contributors. RegMean [13] pioneered a method for merging models trained on diverse datasets for identical tasks, enhancing overall performance. This method relies on a validation set to derive a closed-form solution to a least squares regression problem, maintaining the merged model's weights in close alignment with the original models. Fisher merging [24] assigns varying degrees of importance to different model weights by utilizing the Fisher Information Matrix. Task Arithmetic [12] views the discrepancy between pre-trained and fine-tuned weights as task vectors, averaging these weights across tasks to forge a combined model.

While these techniques show promise for token-based input and output scenarios in language models, they encounter limitations when input modalities shift. Direct application of these methods becomes problematic with changes in input modalities, explaining their rare utilization in multi-task conditioning for vision tasks. Specifically, in

structural conditioning tasks where features vary significantly across modalities, simplistic averaging mechanisms like task merging prove inadequate, often resulting in the generation of multiple undesired artefacts in multi-task conditioned images. Recent efforts, such as GitRebasin [1], have explored fusing models trained on distinct tasks by aligning their weights. However, subsequent analyses, including work by Repair [14], have highlighted the emergence of a 'feature forgetting' issue, where signals from certain tasks weaken in the model's deeper layers. This challenge renders such strategies ineffective for integrating diffusion model weights across modalities for multi-modal generation, as it risks neglecting some conditions.

Therefore, there exists a pressing need for a merging strategy capable of facilitating out-of-domain fusion for diffusion models. A potential solution is a framework proposed by Zipit [38], which maintains two parallel sets of model weights and merges them based on input features. Inspired by this approach, our work seeks to implement a similar strategy for effective model integration.

3 Proposed Method

3.1 Unlocking Incremental Task Addition

ControlNet [49] performs structural conditioning by using dedicated layers designed to handle different types of inputs, enabling it to process diverse input conditions. Once these layers process the inputs, they are fused with the features from the first layer of the diffusion-UNet model. Therefore, we can conclude that, during the training process of ControlNet, the weights of the first layer of different control module input layers, like controllers designed for poses and depth, align the modality to the latent space of a natural image. Hence, the signal input to the blocks is converted to a domain similar to that of the input of the diffusion UNet. Moreover, different modality ControlNets, like depth, sketch, and segmentation mask, handle distinct input modalities. Therefore, it is a must that these input processing blocks remain separate. We walk the readers through the strategy of combining different modalities in the rest of the section. Before proposing a technique combining features of two different models, we first address two crucial questions: (a) When utilizing two different input modalities and the same backbone architecture, how do we align the feature locations across the models processing these inputs for fusion? (b) Given the presence of two different aligned features, which is the most relevant feature vector required to enforce the condition at that particular feature location?

3.2 Which Features are Aligned in Conditioned Diffusion Models?

To address the problem of feature fusion-based schemes for multi-modal generation, a naive solution would be simply averaging the features of every layer of diffusion UNet. However, when the features are not aligned, such a fusion scheme leads to ineffective conditioning. We point out such an example in Section 3. To overcome this, we leverage the model structure of controlled diffusion models. As seen in Figure 4, intermediate layer outputs get added to specific layers of the T2I model. Hence, we make a proposition:

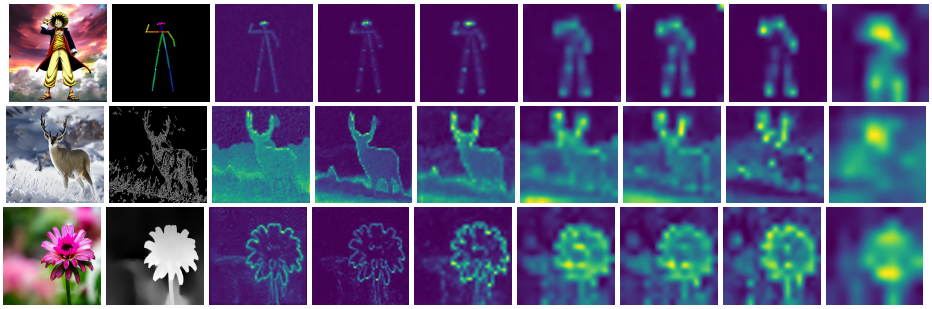


Fig. 3: Variance Maps across channels for intermediate features of ControlNet for different modalities. As we can see the variance map has high values where the condition is present and has low values for locations where the condition is absent.

Observation 1 *Features from different modules that get added to the same spatial location in Stable Diffusion are aligned.*

3.3 Expressiveness of Control Modules

The task of choosing the most suitable feature from two at the same spatial location for fusion poses a significant challenge. To effectively merge intermediate features from multiple models, employing a metric that quantifies the strength or expressiveness of the condition at that specific location becomes essential. For instance, as illustrated in Figure 1, where multiple conditions are showcased, the network encounters the challenge of prioritizing the most expressive condition at each spatial point. Consider a scenario with a sketch map and a segmentation map, as shown in Figure 1(b); the network needs to identify and prioritize the condition that provides the most significant information for that specific location. Our investigations led us to a straightforward yet effective observation.

Observation 2 *The expressiveness of a condition provided to the diffusion model can be quantified using the variance maps from the model's intermediate layers.*

To demonstrate this, we examine variance maps of different control signals originating from an identical image, noise, and text prompt, as depicted in Figure 3. These maps reveal higher activation in areas pertinent to the specific condition, suggesting that variance maps can serve as a measure of conditioning strength needed for each spatial area. Therefore, when comparing two models that process the same input noise but under different conditions, the significance of each condition at every spatial point can be approximated by the activation level of the variance map at that location. This approach provides a reliable means to gauge the relative importance of each condition, enabling more informed feature selection for fusion.

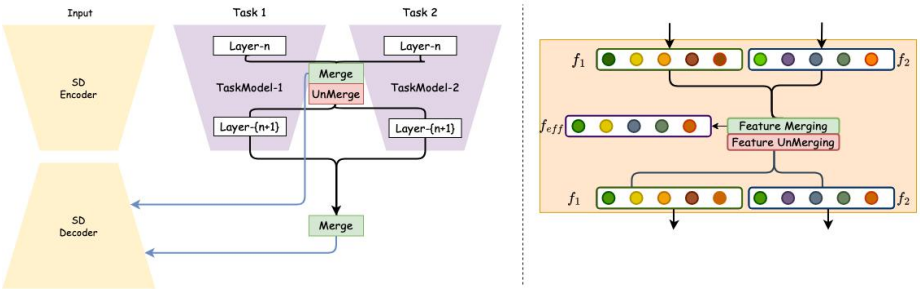


Fig. 4: An illustration of MaxFusion. During the progression through Stable Diffusion, intermediate features are consolidated at each location, and the resulting merged features are subsequently transmitted to the diffusion module. Please note the colours during the merging and unmerging operation. Similar features are not unmerged and are passed as such to the next network layer. Non-similar intermediate outputs are passed without any change.

3.4 MaxFusion

This section delves into the fusion mechanism among the intermediate outputs from various task models passed to the T2I model and the subsequent layer. The InterModel fusion occurs through a merging process detailed in the following paragraphs.

Let us begin with a straightforward scenario: fusing two independent modalities. Let the ControlNet blocks handle different modalities M_1, M_2. A basic approach for combining these modalities is naive averaging, represented as:

$$f_{avg} = \frac{f_1 + f_2}{2} \tag{1}$$

Where f_1 and f_2 are intermediate feature outputs across the models. However, such a simplistic fusion method poses a challenge. Consider an image where the left half contains only depth details while the right half features an edge. A naive averaging scheme would dilute the significance of each feature by averaging them equally. An ideal scheme would adjust spatially, giving more weight to the modality input value at a given location (j, k). In cases of overlapping conditions, such as when conditions denote the same object and spatial location, as illustrated in the robot example in Figure 1, it is desirable to incorporate characteristics from both modalities during merging. To achieve this, we assess the correlation of features at each spatial location. For highly correlated features, we adopt a weighted summation approach formulated as:

$$f_{eff}^{(j,k)} = \frac{f_1^{(j,k)} + f_2^{(j,k)}}{2}, \text{if } \rho^{(j,k)} > \delta \tag{2}$$

where δ is a predefined threshold and $\rho^{(j,k)}$ represents the correlation value between $f_1^{(j,k)}$ and $f_2^{(j,k)}$ at each spatial location and is defined as:

$$\rho^{(j,k)} = \frac{f_1^{(j,k)} \cdot f_2^{(j,k)}}{|f_1^{(j,k)}| \cdot |f_2^{(j,k)}|} \tag{3}$$

If the correlation falls below the threshold, we prioritize the modality with the most spatial information for Stable Diffusion. The feature variance at the location determines this prioritization. To ensure fairness across modalities with differing absolute variance values, we propose a relative standard deviation measure by normalizing standard deviation values across spatial locations:

$$\hat{\sigma}_i^{(j,k)} = \frac{\sigma_i^{(j,k)}}{\sum^{(j,k)} \sigma_i^{(j,k)}} \tag{4}$$

Hence, when the correlation is below the threshold, we define the aggregated feature as the one with the highest channel standard deviation:

$$f_{eff}^{(j,k)} = f_i^{(j,k)}, \max_i(\hat{\sigma}_i^{(j,k)}), \text{if } \rho^{(j,k)} < \delta \tag{5}$$

This formulation ensures automatic selection of the more relevant spatial condition, even in conflicting cases, providing effective conditioning.

3.5 InterModel Unmerging

In this section, we discuss the inter-model fusion process that proceeds to the next layer of the model. During the InterModel merging stage, we evaluate whether the correlation value exceeds the predefined threshold. For each spatial point not meeting this criterion, we verify if the feature being passed at that spatial location corresponds to the same index it belonged to and has the maximum variance. If the element possesses the maximum variance, we pass it as such; otherwise, we rescale the feature vector to have the same standard deviation before merging.

$$f_i^{(j,k)} = \begin{cases} f_i^{(j,k)} & i = \max_i(\hat{\sigma}_i^{(j,k)}) \\ \frac{\sigma_i^{(j,k)}}{\sigma_{max}^{(j,k)}} f_{max}^{(j,k)} & otherwise \end{cases} \tag{6}$$

where,

$$f_{max}^{(j,k)} = f_i^{(j,k)}; i = \max_i(\sigma_i^{(j,k)}) \tag{7}$$

Please note that renormalization of variance in Equation (6) reactivates regions that are otherwise diminished during the estimation of f_{eff}. This effect of vanishing feature strength is negated and rescaled to improve the performance. We detail the overall algorithm in Algorithm 1, where SD denotes Stable Diffusion and $l1, l2$ are model layers, and we name the input modalities as $c1, c2$.

4 Experiments and Results

4.1 Implementation Details

All experiments are conducted on a single NVIDIA A5000 GPU utilizing off-the-shelf models. We employ Stable Diffusion-v1.5 [34] as the T2I model. For task models,

Algorithm 1 MaxFusion for scaling two modalities

Input: Model 1 layers l_1, Model 2 layers l_2 , threshold δ, input condition c_1, input condition c_2

for $l = 1, \ldots, L$ **do**
  if $l = 1$ **then**
    $f_1 = l_1(c_1)$
    $f_2 = l_2(c_2)$
  else
    if $\rho(f_1^{(j,k)}, f_2^{(j,k)}) \geq \delta$ **then**
      $f_{eff}^{(j,k)} = \frac{f_1^{(j,k)} + f_2^{(j,k)}}{2}$
    else
      $f_{eff}^{(j,k)} = f_i^{(j,k)}, \max_i(\hat{\sigma}_i^{(j,k)})$
    end if
    SD decoder $\leftarrow feff$
    if $\rho(f_1^{(j,k)}, f_2^{(j,k)}) \geq \delta$ **then**
      $fi^{(j,k)} = f_{eff}^{(j,k)}; i = 1, 2$
    else
      $$f_i^{(j,k)} = \begin{cases} f_i^{(j,k)} & i = \max_i(\hat{\sigma}_i^{(j,k)}) \\ \frac{\sigma_i^{(j,k)}}{\sigma_{max}^{(j,k)}} f_{max}^{(j,k)} & otherwise \end{cases}$$
    end if
  end if
end for return f_{eff}

we use ControlNet [49] and T2IAdapter [26] modules trained separately for different tasks. Throughout our experiments, we maintain a consistent diffusion duration of 50 steps. Sampling is facilitated using the UniPC scheduler [51]. In our experiments, we set the correlation threshold at a fixed value of 0.7. Due to the absence of public datasets tailored for evaluating multi-modal generation with multiple modality inputs, we curate a synthetic dataset leveraging the COCO dataset [19]. This dataset comprises modalities including depth maps, segmentation masks, HED, and canny edges. Similar to the approach in ControlNet [49], we utilize MiDAS [16] for depth map estimation, PiDiNet [39], the canny edge detector, and UPERNET [43] for generating maps of individual modalities. We curate a subset of 2000 images from the COCO validation set for experimentation purposes. We utilize BLIP [18] to acquire corresponding captions and perform evaluations under dual conditions. For evaluations, we leverage variants of existing networks and publicly available single-modality networks.

4.2 Qualitative Evaluations

We present results for multi-modal conditions in Figure 5 and 6. Here, we compare with uni-modal networks SPADE [30] and PITI [41]. For multi-modal conditioning, we implement ControlNet and T2IAdapter with different conditioning scales for the various input tasks. On the left, we display the included modalities. As observed, the single-modality segmentation model SPADE [30] fails to capture edges clearly, as it is trained solely with segmentation maps. Similarly, PITI often struggles with image quality in most cases, as its training data may not adequately represent complex scenar-

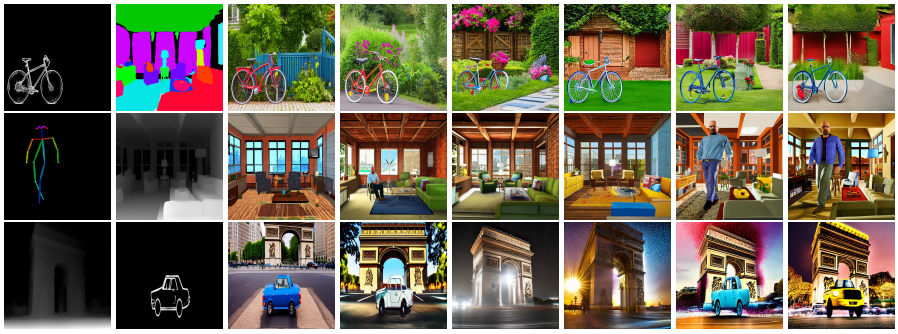


Fig. 5: Qualitative comparisons for contradictory conditions from different modalities. Figure illustrates the case of different conditions where out method shows a clear advantage over existing approaches. The text prompts used are *(1) "A bicycle in a garden." (2): "Walter White in a living room, GTA style." (3) "A car in front of arc de triomphe"*

Method	Pose, Seg			Pose, Depth		
	FID (↓)	MSE-P (↓)	MSE-S (↓)	FID (↓)	MSE-P (↓)	MSE-D (↓)
T2I-Adapter	55.61	0.0140	0.1338	51.47	0.0134	0.0603
ControlNet	63.46	0.0142	0.1276	48.18	0.0137	0.0357
Ours	46.72	0.0135	0.1080	47.37	0.0119	0.0410

Table 1: Metrics for contradictory conditions on COCO

ios like "A banana on a chair". The multi-modal networks T2I-Adapter and ControlNet generally perform well across most cases, yet they may falter in generating intricate details in challenging scenarios. Please refer to Figure 6 for examples from the top two rows. On a broader scale, multi-controlnet performs quite well but struggles to preserve key details when multiple conditions are present at the same location, as depicted in the provided examples. Our proposed method successfully transfers details and performs well in challenging cases while maintaining consistency with the text prompt. Additional illustrative examples are available in the supplementary material.

4.3 Quantitative Evaluations:

Quantitative Analysis for Contradictory Conditions on COCO: Since existing datasets do not encompass contradictory inputs in COCO, and general strategies for deriving spatial conditioning maps do not provide such datasets, we created a synthetic dataset for evaluation purposes. We selected all COCO images containing humans and derived their pose maps. Additionally, we included an extra pose map sourced from another set of images obtained from the internet. Consequently, the pose maps now incorporate additional spatial features, while other spatial features express the scene. The quantitative analysis is presented in Table 1.

Quantitative Analysis for Complementary Conditions on COCO: For experiments involving segmentation maps, we employ SPADE [30] as the single-modality model. For experiments involving sketches, we utilize PITI [41] as the single-modality model. As for the choice of multi-modal models, we establish the baseline using T2IAdapter [26] with naive averaging, simultaneously feeding in multi-modality inputs. We employ

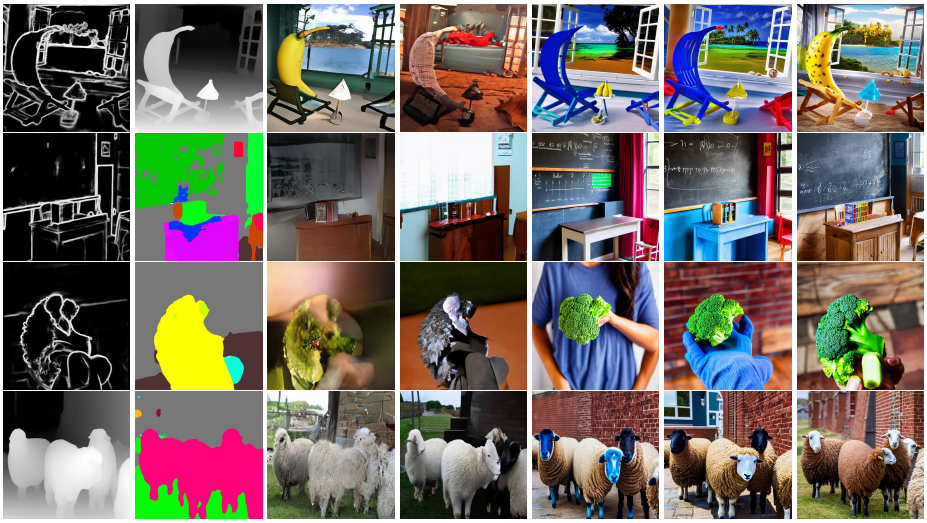


Fig. 6: Qualitative comparisons for complimentary conditions in COCO dataset. For the unimodal models we condition with its corresponding conditions. For multimodal models, we condition using both the Tasks. The text prompts used for conditioning are: from top (1): *"A banana sitting on a chair and looking at the beach." (2) "A brocolli in a hand." (3) "A blackboard and table in a classroom" (4) "A couple of sheep standing"*

Method	Modality	FID(↓)	NIQE(↓)	Clip(↑)	Depth(↓)	Seg(↓)
SPADE [30]	*Uni*	41.05	4.95	0.2813	0.0320	0.0923
PITI-Mask [41]	*Uni*	35.52	6.20	0.2816	0.0269	0.0909
T2I-Adapter [26]	*Multimodal*	43.24	5.34	0.2955	0.0554	0.1387
Ours	*Multimodal*	38.84	5.34	0.2954	0.0296	0.1107

Method	Modality	FID(↓)	NIQE(↓)	Clip(↑)	Depth(↓)	HED(↓)
PITI-Sketch [21]	*Uni*	38.50	6.22	0.2669	0.0298	0.0417
T2I-Adapter [26]	*Multimodal*	30.63	3.90	0.2964	0.0417	0.1055
Ours	*Multimodal*	29.71	2.86	0.2879	0.0191	0.0735

Table 2: Quantitative comparisons for $\{Seg, Depth \rightarrow Image\}$ **and**$\{Hed, Depth \rightarrow Image\}$ respectively .

NIQE [25] and CLIP consistency [33] as evaluation metrics in experiments combining Depth-Segmentation. Moreover, we derive the conditions from the generated images and evaluate them using Mean Square Error against their ground truth values to ensure conditioning consistency. Observations reveal that Unimodal networks perform admirably in terms of conditioning consistency. However, they fall short in terms of text consistency and generated image quality. On the other hand, multi-modal networks exhibit superior performance and yield overall better metrics in terms of consistency. Notably, our approach demonstrates better CLIP scores than other evaluation methods, ensuring enhanced consistency across text prompts.

5 Discussions

Why is using the proposed variance scheme advantageous rather than naively adding the spatial conditions at the control input? During our experiments, we observed that when both conditions are complementary, i.e., when we provide more than one condition at the same spatial location, naive averaging of the features works well

Method	Modality	FID(↓)	NIQE(↓)	Clip(↑)	Seg(↓)	HED(↓)
SPADE [30]	*Uni*	41.05	4.95	0.2813	0.0923	0.0772
PITI-Mask [41]	*Uni*	35.52	6.20	0.2816	0.0909	0.0844
PITI-Sketch [41]	*Uni*	38.50	6.22	0.2669	0.1044	0.0417
T2I-Adapter [26]	*Multimodal*	40.14	3.94	0.2922	0.1506	0.1201
Ours	*Multimodal*	37.21	2.84	0.2941	0.0985	0.0774

Table 3: Qualitative comparisons for $\{Hed, Seg \rightarrow Image\}$. Quantitative results for different models for multimodal generation. Please note that the multimodal models have additional text conditioning capability

for most cases. This is evident in Figure 6, where we can observe satisfactory performance for the naive averaging scheme. However, when the conditions contradict each other, as shown in Figure 5, the naive averaging scheme fails to effectively incorporate details from both modalities into the scene. Our approach offers a clear advantage over existing approaches in this scenario, opening up possibilities for zero-shot generation by deriving details of different objects from different input types.

Going beyond spatial conditioning: Recently, a trained version of Multi-T2I-Adapter has shown the capability to include style conditioning as an additional modality. To show that MaxFusion also utilizes style as a modality, we present the corresponding results with different conditioning techniques in Figure 7. Additional results are provided in the supplementary material.

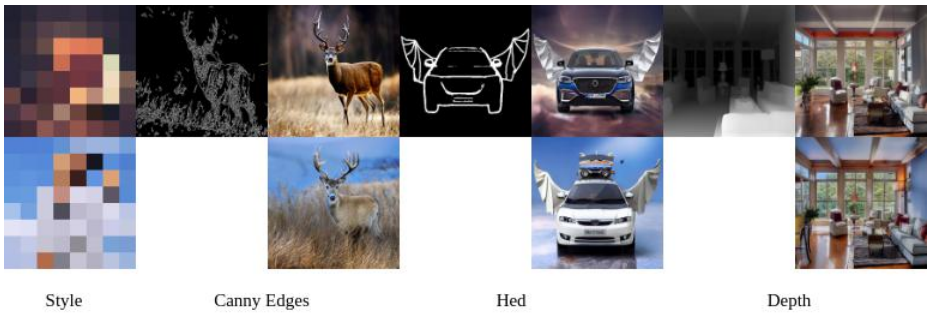


Fig. 7: Qualitative results with style as a modality

Scalability beyond two modalities: To show that MaxFusion extends beyond two modalities, we show that one can change Maxfusion to combine two modalities first and add the next one on top of it and continue this process until all tasks are added. Let us assume we have three tasks: Task1, Task2, Task3. One can first consider two at a time and add the third task on the obtained intermediate feature. The incremental addition scheme is the same for intra-modal and inter-modal fusion. For extending to N tasks, one can follow the same procedure, incrementally adding each task to the previous. We show illustrative examples for the same in Figure 8.



Fig. 8: Qualitative results for extending to 3 modalities.

Method	{Seg, Depth, HED}			
	FID(↓)	MSE-S(↓)	MSE-D(↓)	MSE-H(↓)
T2I-Adapter	51.87	0.1354	0.0448	0.1046
ControlNet	50.47	0.1019	0.0362	0.0758
Ours	47.48	0.1000	0.0204	0.0692

Table 4: Scalability metrics, scaling to 3 modalities.

6 Ablation Studies

Effect of correlation Value for Multi-Modal Generation: We present the results of varying the correlation threshold from 0 to 1 in Figure 9. When the correlation threshold is 0, naive averaging of features occurs, leading to reduced depth and edges conditioning strength and resulting in inconsistent depth effects in the scene. Specifically, at $\delta = 0$, the image lacks 3D depth-based effects in regions where the teddy bear is situated. Conversely, more realistic images are generated when the correlation threshold $\delta = 1$. Varying the correlation threshold impacts sample quality, with a threshold above $\delta > 0.5$ yielding realistic images. We set the value of δ to 0.7 to mitigate very high-frequency out-of-distribution components in the image. We observed that $\delta = 0.75$ works well for most cases.



Fig. 9: Ablation study corresponding to different correlation thresholds. Please note the level of 3D effect included to the scene as the threshold increases. The prompt used for captioning is *Background:"A mountain". Foreground : "A teddy bear".*

Effect of Multi-modal Conditioning: We present the results on single conditioning and multi-conditioning in Figure 10. When conditioned with edges alone, the image appears planar and visually unpleasing. Similarly, an additional wing is generated at a specific location when conditioned only on depth. In cases like this, more aesthetically pleasing, high-quality images with fine-grained details from 2D edge conditions and 3D depth effects can be generated using multi-modal generation.



Fig. 10: Ablation study drawing contrast between unimodal-models and multimodal models. The prompt used for captioning is *"A bird"*.

7 Limitations, Future Works, and Potential Impacts

Our model inherently inherits the limitations of Stable Diffusion in generating human hands and faces. Additionally, we observed discrepancies in results when utilizing semantic maps, sometimes differing from the conditioned semantic map. However, this limitation stems from the training approach of ControlNet and T2I-Adapter. As state-of-the-art semantic segmentation networks cannot provide accurate masks for all open-world images, this limitation is also inherent in segmentation-conditioned models. We provide examples in the paper and supplementary material to elucidate this issue. Furthermore, while our work can generate various spatial inputs, it holds potential societal impacts. Also, one other limitation of our method is that as the number of conditions increases, a trade-off between the conditioning and the sampling fidelity arises. We will tackle this issue in future works by developing a much more robust algorithm. In addition, there is a slight drop in performance when the number of modalities is increased. One potential way to improve this is by considering all modalities simultaneously and re-deriving the equations of Maxfusion for N modalities. Like other generative models, our model also has potential societal impacts. Using such models is at the user's discretion, and care must be taken while using the model.

8 Conclusion

In this paper, we discover that the expressiveness of conditioning features is present in the variance maps of intermediate layers of diffusion models. Utilizing this as prior information, we introduced a novel training-free strategy to scale up conditional diffusion models for multi-modal generation using off-the-shelf models trained on individual tasks separately. For multi-modal fusion, we proposed a new fusion paradigm based on variance across channels for different feature layers. The proposed fusion approach is lightweight and can be seamlessly integrated with existing models. This new fusion scheme empowers the model to perform zero-shot generation, which is otherwise not achievable with a single model. We further showed that the proposed method works beyond spatial conditioning and with style conditioning. Moreover, Maxfusion can be seamlessly integrated into more than two conditioning modalities. We conducted experiments with four different spatial conditioning modalities to demonstrate the effectiveness of our method.

Acknowledgements Vishal M Patel was supported by NSF CAREER award 2045489

References

1. Ainsworth, S.K., Hayase, J., Srinivasa, S.: Git re-basin: Merging models modulo permutation symmetries. arXiv preprint arXiv:2209.04836 (2022)
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
4. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
5. Du, Y., Durkan, C., Strudel, R., Tenenbaum, J.B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., Grathwohl, W.S.: Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In: International Conference on Machine Learning. pp. 8489–8510. PMLR (2023)
6. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. arXiv preprint arXiv:2212.05032 (2022)
7. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22930–22941 (2023)
8. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
10. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)
11. Huang, Z., Chan, K.C., Jiang, Y., Liu, Z.: Collaborative diffusion for multi-modal face generation and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6080–6090 (2023)
12. Ilharco, G., Ribeiro, M.T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., Farhadi, A.: Editing models with task arithmetic. arXiv preprint arXiv:2212.04089 (2022)
13. Jin, X., Ren, X., Preotiuc-Pietro, D., Cheng, P.: Dataless knowledge fusion by merging weights of language models. arXiv preprint arXiv:2212.09849 (2022)
14. Jordan, K., Sedghi, H., Saukh, O., Entezari, R., Neyshabur, B.: Repair: Renormalizing permuted activations for interpolation repair. arXiv preprint arXiv:2211.08403 (2022)
15. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
16. Kopf, J., Rong, X., Huang, J.B.: Robust consistent video depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1611–1621 (2021)
17. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)

18. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
20. Lingenfelter, B., Davis, S.R., Hand, E.M.: A quantitative analysis of labeling issues in the celeba dataset. In: International Symposium on Visual Computing. pp. 129–141. Springer (2022)
21. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
22. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)
23. Lu, H., Tunanyan, H., Wang, K., Navasardyan, S., Wang, Z., Shi, H.: Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14267–14276 (2023)
24. Matena, M.S., Raffel, C.A.: Merging models with fisher-weighted averaging. Advances in Neural Information Processing Systems **35**, 17703–17716 (2022)
25. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
26. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
27. Nair, N.G., Bandara, W.G.C., Patel, V.M.: Unite and conquer: Cross dataset multimodal synthesis using diffusion models. arXiv preprint arXiv:2212.00793 (2022)
28. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
29. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
30. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
31. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
32. Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., Wang, H., Niebles, J.C., Xiong, C., Savarese, S., et al.: Unicontrol: A unified diffusion model for controllable visual generation in the wild. arXiv preprint arXiv:2305.11147 (2023)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
35. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242 (2022)

36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
37. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
38. Stoica, G., Bolya, D., Bjorner, J., Hearn, T., Hoffman, J.: Zipit! merging models from different tasks without training. arXiv preprint arXiv:2305.03053 (2023)
39. Su, Z., Liu, W., Yu, Z., Hu, D., Liao, Q., Tian, Q., Pietikäinen, M., Liu, L.: Pixel difference networks for efficient edge detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5117–5127 (2021)
40. Sung, Y.L., Li, L., Lin, K., Gan, Z., Bansal, M., Wang, L.: An empirical study of multimodal model merging. arXiv preprint arXiv:2304.14933 (2023)
41. Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F.: Pretraining is all you need for image-to-image translation. arXiv preprint arXiv:2205.12952 (2022)
42. Weinbach, S., Bellagente, M., Eichenberg, C., Dai, A., Baldock, R., Nanda, S., Deiseroth, B., Oostermeijer, K., Teufel, H., Cruz-Salinas, A.F.: M-vader: A model for diffusion with multimodal context. arXiv preprint arXiv:2212.02936 (2022)
43. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
44. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7754–7765 (2023)
45. Yadav, P., Tam, D., Choshen, L., Raffel, C., Bansal, M.: Ties-merging: Resolving interference when merging models. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
46. Yang, E., Wang, Z., Shen, L., Liu, S., Guo, G., Wang, X., Tao, D.: Adamerging: Adaptive model merging for multi-task learning. arXiv preprint arXiv:2310.02575 (2023)
47. Zeng, Y., Lin, Z., Zhang, J., Liu, Q., Collomosse, J., Kuen, J., Patel, V.M.: Scenecomposer: Any-level semantic image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22468–22478 (2023)
48. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. arXiv preprint arXiv:2306.10012 (2023)
49. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
50. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10146–10156 (2023)
51. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. arXiv preprint arXiv:2302.04867 (2023)