Cocktail : Mixing Multi-Modality Controls for Text-Conditional Image Generation

Minghui Hu[†], Jianbin Zheng[⋆,‡], Daqing Liu[‡], Chuanxia Zheng[§], Chaoyue Wang[‡], Dacheng Tao[‡], and Tat-Jen Cham[†]

[†]*Nanyang Technological University,* [⋆]*South China University of Technology,* [§]*University of Oxford,* [‡]*JD Explore Academy*

Abstract

Text-conditional diffusion models are able to generate high-fidelity images with diverse contents. However, linguistic representations frequently exhibit ambiguous descriptions of the envisioned objective imagery, requiring the incorporation of additional control signals to bolster the efficacy of text-guided diffusion models. In this work, we propose Cocktail, a pipeline to mix various modalities into one embedding, amalgamated with a generalized ControlNet (gControlNet), a controllable normalisation (ControlNorm), and a spatial guidance sampling method, to actualize multi-modal and spatially-refined control for text-conditional diffusion models. Specifically, we introduce a hyper-network gControlNet, dedicated to the alignment and infusion of the control signals from disparate modalities into the pre-trained diffusion model. gControlNet is capable of accepting flexible modality signals, encompassing the simultaneous reception of any combination of modality signals, or the supplementary fusion of multiple modality signals. The control signals are then fused and injected into the backbone model according to our proposed ControlNorm. Furthermore, our advanced spatial guidance sampling methodology proficiently incorporates the control signal into the designated region, thereby circumventing the manifestation of undesired objects within the generated image. We demonstrate the results of our method in controlling various modalities, proving high-quality synthesis and fidelity to multiple external signals. The codes will be released at https://mhh0318.github.io/cocktail/.

1 Introduction

Text-conditional diffusion models [40, 38, 41, 33] have actualized the capacity for high-quality generative capabilities. These models facilitate the generation of an array of high-calibre images through the utilization of concise textual prompts. However, linguistic representations pose inherent challenges in accurately encapsulating the precise imagery anticipated by the user, owing to the potential ambiguities and subjective interpretations of verbal descriptions within the context of visual synthesis. Moreover, minor alterations to the textual prompts also yield distinct visual outputs, underscoring the absence of refined control over the generative process.

Modifying the prior is a series of existing solutions for multi-modal control, *e.g.*, the control of the entire prior space [38, 47, 22, 21, 56] as demonstrated in Fig. 1(a). These approaches centered on the whole prior lack the capacity for localised image modifications and the preservation of background elements. Moreover, these models typically require training from scratch, which demands a substantial amount of resources.

Preprint. Under review.

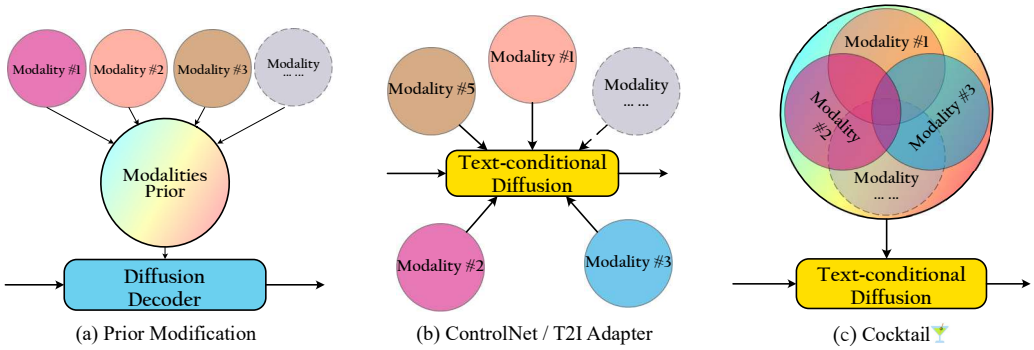

Figure 1: **Comparison of various control methods.** Our approach requires only one generalized model, unlike previous that needed multiple models for multiple modalities.

In response to the methods dealing with latent representation [28, 4, 26], an additional lightweight hyper-network is introduced in [52, 32], which is designed to encode external control signals into latent vectors and subsequently inject them directly into the backbone network. Such methods effectively handle control over the single additional modality; however, they exhibit limitations when confronted with multiple modalities, as shown in Fig 1(b). One issue is that each modality necessitates a unique network, leading to a computational overhead that escalates proportionally with the increase in the number of modalities. Furthermore, the impact of additive coefficients between different modes on the final imaging outcomes warrants consideration. The inherent imbalance among superimposed modes makes these additive coefficients a pivotal factor in determining the ultimate synthesized output. An additional challenge emerges during the sampling process when the model conducts an initial inference devoid of control signal injection. This preliminary inference step could result in object placement that potentially contradicts the control signals.

In this paper, we propose a novel pipeline, as shown in Fig. 1(c), termed Cocktail, which accomplishes multi-modality control using the text-conditional diffusion model as the foundational spirit. It encompasses three main components: a hyper-network capable of accommodating multi-modal input, a conditional normalisation method to mix the control features, and a sampling strategy designed to facilitate precise control over the generative process. As shown in Fig. 2, our method is proficient in generating images that meet all input conditions or any arbitrary subset thereof, utilizing one single model.

We initially trained a branched network named gControlNet, which accepts multiple modalities. Upon training, the branched network can simultaneously accept arbitrary combinations of existing modalities. When multiple modalities coexist within the same region, the model is capable of automatically fusing input from different modalities, balancing the disparities between them. To leverage the features of the gControlNet, we further proposed controllable normalisation (ControlNorm). We can achieve better representation of control signals in terms of semantic and spatial aspects according to the decoupling provided by ControlNorm.

Moreover, we introduce a spatial guidance sampling method in order to facilitate spatial generation under the purview of multi-modal control signals. Specifically, we employ distinct textual prompts to differentiate entities from the background, incorporating entities into the background via a prompt editing approach. Our devised sampling approach demonstrates efficacy in circumventing the generation of undesired entities.

The contribution of the proposed pipeline is threefold:

- We introduced the Generalized ControlNet (gControlNet), a branched network capable of adaptively integrating multi-modal information, effectively addressing issues stemming from imbalances between modalities;
- We proposed the Controllable Normalistion (ControlNorm) to optimize the utilization of information within branched networks to yield more effective outcomes;

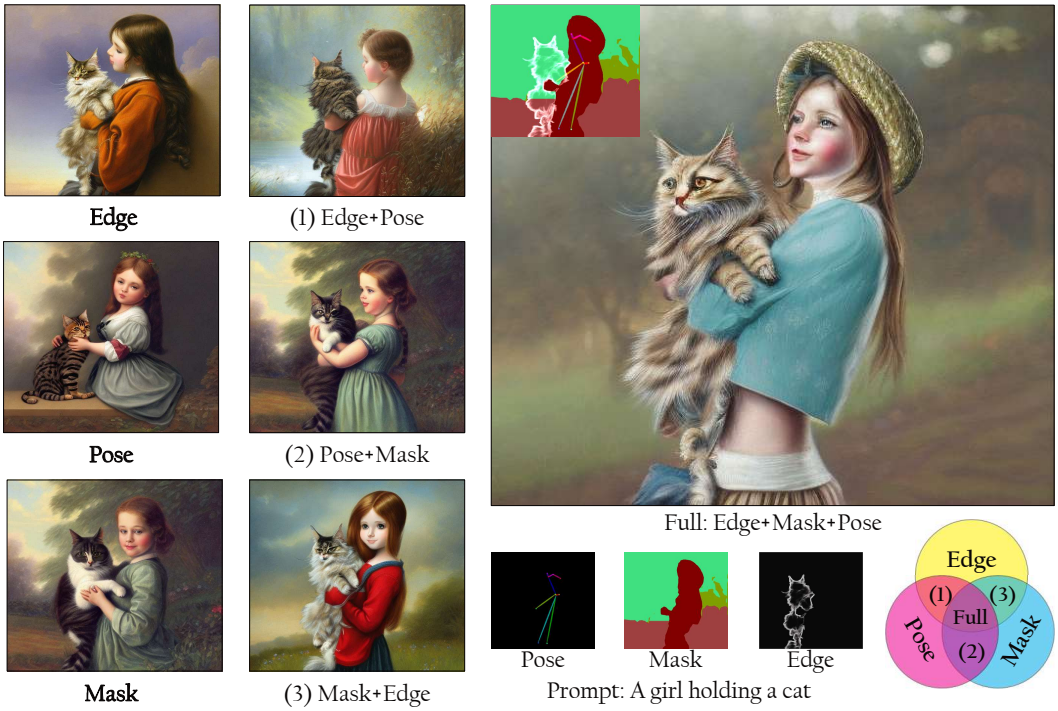

Figure 2: **Examples of our model with the same prompt.** Given a text prompt along with various modality signals, our approach is able to synthesize images that satisfy *all input conditions* or *any arbitrary subset* of these conditions using *a single model*. The prompt is: *A girl holding a cat.*

- We introduced a spatial guidance sampling method based on the operation within the attention map to generate relevant information tailored to regional contexts, preventing the inclusion of undesired objects outside the specified regions.

2 Related Work

2.1 Text-conditional Diffusion Models

Diffusion models [19, 45] has achieved great success in the area of text-to-image synthesis [33, 38, 40, 3, 14]. To reduce computational cost, diffusion models typically function within the latent space [40] or produce low-resolution images that are later improved through super-resolution models [38, 3]. Fast sampling methods have also successfully reduced the number of generation steps required by diffusion models from hundreds down to merely a few [44, 31, 24, 29, 11]. The provision of classifier guidance during the sampling process can also significantly impact the outcomes, leading to substantial improvements in the results [9]. In addition to the widely used classifier-free guidance [18], other types of guidance are also worth exploring [54, 13].

2.2 HyperNetworks for Pre-trained Models

Training a diffusion model is highly resource-intensive and environmentally unfriendly [40]. Fine-tuning such a model can also be challenging due to the vast number of parameters involved [1]. Therefore, introducing an additional tiny branched network to bias the output of the original network is a more reasonable choice [10, 2, 8]. Similar ideas have also proven to be effective in diffusion models: Hypernet [15] and LoRA [20] models are capable of altering the original diffusion model's sampling distribution by training a small branched network, which has become the most popular branched network in the current research community. More similar to our work is ControlNet [52], which are subsequently combined with different layers in the denoising U-Net to provide support for various task-specific guidance, *with only one modality for each model.* In contrast, our cocktail endows the ControlNet with multitasking capabilities.

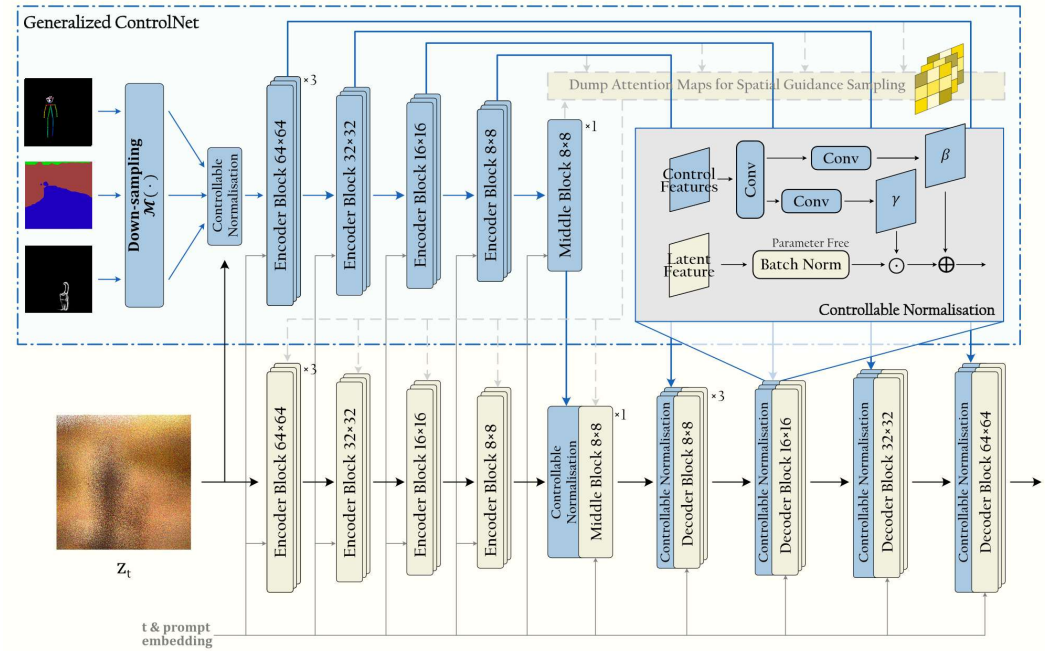

Figure 3: **The network architecture** of Generalized ControlNet (gControlNet) with Controllable Normalisation (ControlNorm). The parameters indicated by the yellow sections are sourced from the pre-trained model and stay constant, while only those in the blue sections are updated during training, with the gradient back-propagated along the blue arrows.

2.3 Conditional Normalisation

Conditional normalisation approaches have been employed across a range of vision tasks, including style transfer [23, 12], conditional generation [35, 7, 14, 55] and image-to-image translation [39]. These techniques involve normalizing layer activations to zero mean and unit deviation, followed by denormalisation through an affine transformation derived from external data. External data can be in multiple formats, such as style images, semantic masks, or category labels.

2.4 Attention Map-based Prompt Tuning

Prompt-to-Prompt [16] is a method that adjusts local or global specifics in text-guided diffusion models by altering the cross-attention maps from source to target image, thereby maintaining spatial layout and geometry. Recently, numerous efforts have been made to improve the outcomes [34, 36]. However, such approaches are limited to synthesized images without an inversion technique. Unlike cross-attention, self-attention focuses on inter-pixel relationships within the same domain [46]. Paint-with-words [3] is another method that allows users to specify the spatial locations of objects by selecting phrases from the text prompt.

3 Methods

In this work, we first propose a more general branched network, gControlNet, which can generate control signals from different modalities using a single network and adaptively weighted fuse them together. Furthermore, we propose ControlNorm to inject the signals from the gControlNet into the diffusion backbone, which solves the imbalanced problem within various modalities. Finally, we propose a spatial guidance sampling method to avoid the presence of extraneous objects in the generation process. By modifying the attention map, this approach effectively incorporates control signals into the backbone network. The whole pipeline is demonstrated in Fig. 3.

3.1 Generalized ControlNet with Controllable Normalisation

Generalized ControlNet ControlNet [52] is a method designed to influence and control the behavior of neural networks by adjusting the input conditions of specific network blocks. Instead of directly modifying the parameters of the primary network, ControlNet employs an auxiliary network to generate feature offsets. These offsets are then combined with different layers in the main network, such as a denoising U-Net, to support various task-specific guidance.

Given a trained backbone network block $\mathcal{F}(\cdot;\boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$, the input feature $\boldsymbol{x}$ can be mapped to $\boldsymbol{y}$. For the branched part, we duplicate the parameter $\boldsymbol{\theta}$ to create a trainable copy $\boldsymbol{\theta}_t$, which is then trained using the supplementary modality. Preserving the original weights helps retain the information stored in the initial model after training on large-scale datasets, which ensures that the quality and diversity of the generated images do not degrade. Mathematically, the output from the trained network block can be expressed as:

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x};\boldsymbol{\theta}) + \mathcal{Z}\left(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}_m);\boldsymbol{\theta}_t)\right) \leftarrow \mathcal{F}(\boldsymbol{x};\boldsymbol{\theta}), \tag{1}$$

where the control signal of a single modality $\boldsymbol{c}_m$ is typically processed to obtain a format identity for $\boldsymbol{x}$, for example, through a zero-initialized convolutional layer, and then added to $\boldsymbol{x}$. $\mathcal{Z}(\cdot)$ represents the zero-initialized layer. It is not only used in the process of handling $\boldsymbol{c}_m$, but also serves to adjust the output of the branched network $\mathcal{F}(\cdot,\boldsymbol{\theta}_t)$.

To accomplish the goals of accepting multiple external modalities as input and balancing signals from different modalities, we have devised a modified framework that adeptly merges these varied sources of information. At the top of our network, we adopt a simple downsampling network $\mathcal{M}(\cdot)$ to convert external conditional signals to the latent space, allowing the conditional signals to be directly injected into the latent space. It is worth noting that $\mathcal{M}(\cdot)$ is versatile and can adapt to different types of external signals. Given k different modalities, the converted conditional features are $\boldsymbol{c}_m^k = \mathcal{M}(C^k)$.

Controllable Normalisation Instead of directly passing the sum of conditional features via a zero-initialized layer to the network block $\mathcal{F}(\cdot;\boldsymbol{\theta}_t)$, *i.e.*, $\hat{\boldsymbol{c}}_m = \mathcal{Z}(\sum_i \boldsymbol{c}^i)$, we instead introduce a *controllable normalisation (ControlNorm)* method, which has an additional layer to generate two sets of learnable parameters, $\boldsymbol{\gamma}(\hat{\boldsymbol{c}}_m)$ and $\boldsymbol{\beta}(\hat{\boldsymbol{c}}_m)$, conditioned on all k modalities. These two sets of parameters are used in the conditional normalisation layer to fuse the external conditional signals and the original signals. Specifically, the input to the trainable block $\mathcal{F}(\cdot;\boldsymbol{\theta}_t)$ becomes:

$$\left(\boldsymbol{I} + \mathcal{Z}(\boldsymbol{\gamma}\left(\hat{\boldsymbol{c}}_m\right))\right) \odot \frac{\boldsymbol{x} - \mu_c(\boldsymbol{x})}{\sigma_c(\boldsymbol{x})} \oplus \mathcal{Z}(\boldsymbol{\beta}(\hat{\boldsymbol{c}}_m)) \leftarrow \boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}_m), \tag{2}$$

where $\mu_c(\boldsymbol{x})$ and $\sigma_c(\boldsymbol{x})$ are the mean and standard deviation of the feature $\boldsymbol{x}$ along the channel c, $\odot$ is Hadamard product, and $\boldsymbol{\gamma}(\hat{\boldsymbol{c}})$ and $\boldsymbol{\beta}(\hat{\boldsymbol{c}})$ are vectors that have the same dimension as $\boldsymbol{x}$. With the help of zero-convolution, we can preserve the identity of x just before the start of the fine-tuning. It is worth noting that in the following layers, the internal feature in the branched network $\boldsymbol{h}$ can be integrated with the latent feature $\boldsymbol{z}$ from the original network in the same way:

$$\left(\boldsymbol{I} + \mathcal{Z}(\boldsymbol{\gamma}\left(\boldsymbol{h}\right))\right) \odot \frac{\boldsymbol{z} - \mu_c(\boldsymbol{z})}{\sigma_c(\boldsymbol{z})} \oplus \mathcal{Z}(\boldsymbol{\beta}(\boldsymbol{h})) \leftarrow \boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}_m), \tag{3}$$

where $\boldsymbol{h}$ is the intermediate features from the original network and $\boldsymbol{z}$ is the intermediate features from the branched network. Specifically, we will have five sets of intermediate features from the Stable Diffusion [40] U-Net backbone and our generalized ControlNet, including four sets of features from encoder blocks and one from the middle block.

In fact, our controllable normalisation is a generalized version of conditional normalisation [35, 23]. After changing the mean and variance calculation dimension and replacing the external signal $\hat{\boldsymbol{c}}$ by a mask image, real image, or class labels, we can derive the various forms of SPADE [35], AdaIN [23], CIN [12] and MoVQ [55]. More interestingly, our controllable normalisation method not only enables the use of external signals as conditions, but also allows intermediate layer signals to act as constraints.

As shown in Fig. 3, only the parameters of the gControlNet require updating. Given an initial latent $\boldsymbol{z}_0$ and a time-step t, the diffusion process progressively introduces noise ϵ to this latent, transforming the original latent into a noisy state $\boldsymbol{z}_t$. The objective of the diffusion model is to estimate the noise

ϵ adding to the $\boldsymbol{x}$. Our proposed gControlNet shares the same objective function as the diffusion model, aiming to predict the noise added at time t. The only distinction lies in the incorporation of multimodal information as conditional signals $\hat{\boldsymbol{c}}_m = \left[\boldsymbol{c}_m^0, \boldsymbol{c}_m^1, \ldots, \boldsymbol{c}_m^k\right]$:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0,t,\boldsymbol{c}_p,\hat{\boldsymbol{c}}_m,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{c}_p, \hat{\boldsymbol{c}}_m\right)\|_2^2\right] \tag{4}$$

3.2 Spatial Guidance Sampling

In order to leverage the control signals from generalized ControlNet and ensure that the generated objects appear within the areas of interest, we proposed a *spatial guidance* sampling method. We mainly focus on editing the cross-attention layers of the U-Net in Stable Diffusion.

Given a noisy latent $\boldsymbol{z}$ at timestep t as the input to the pretrained U-Net with n blocks $\mathcal{F}^{(n)}(\cdot;\boldsymbol{\theta}^{(n)})$ at different resolution, the latent vector $\boldsymbol{z}$ will be down-sampled to various dimensional $\boldsymbol{z}^{(n)}$ that map to the corresponding block $\mathcal{F}^{(n)}$. The cross-attention maps $\boldsymbol{A}^{(n)} \in \mathbb{R}^{(N_i,N_t)}$ in each block, parameterized by $\boldsymbol{\theta}^{(n)}$, are associated with the linear projection of latent feature $Q^{(n)} = f_Q(\boldsymbol{\phi}(\boldsymbol{z}^{(n)}))$ and the prompt $K = f_K(\boldsymbol{c}^{\text{text}})$:

$$A^{(n)}_{ij|\boldsymbol{\theta}^{(n)}} = \frac{\exp\langle Q_i^{(n)}, K_j\rangle}{\sum_{k=1}\exp\langle Q_i^{(n)}, K_k\rangle}, \tag{5}$$

where $A_{ij}^{(n)}$ represents the attentional strength between the j-th prompt token (among N_t tokens) and the i-th latent feature (among N_i features). Intuitively, for each prompt token K_j, there exists a corresponding latent feature map $A_j^{(n)}$ with spatial information. Moreover, the corresponding feature map will contain spatial and shape information for the associated object. However, it is not only the object-describing tokens that contain information about the object's location and shape, some connecting words and padding tokens also convey spatial information for the overall scene [6, 42].

We apply a masking strategy to the corresponding attention maps. In detail, we construct two sets of attention masks $M^{\text{pos}(n)}$ and $M^{\text{neg}(n)} \in \mathbb{R}^{(N_i,N_t)}$. Each column $M_j^{\text{pos}(n)}$ and $M_j^{\text{neg}(n)}$ is a flattened alpha mask, which is determined by the visibility of the corresponding text token K_j. The values of $M_{ij}^{\text{pos}(n)}$ and $M_{ij}^{\text{neg}(n)}$ are determined based on the relationship between image token Q_i and text token K_j. Specifically, if image token Q_i corresponds to a region of the image that should be influenced by text token K_j, $M_{ij}^{\text{pos}(n)}$ is assigned the value of 1. On the other hand, if image token Q_i corresponds to a region of the image that should not be influenced by text token K_j, $M_{ij}^{\text{neg}(n)}$ is set to 1. It worth noting that we consider the feature maps corresponding to most words as negative in order to avoid generating undesired objects.The mask components $M^{\text{pos}(n)}$ and $M^{\text{neg}(n)}$ are incorporated into the cross-attention computation process:

$$\tilde{A}^{(n)}_{ij|\boldsymbol{\theta}^{(n)}} = \frac{\exp\langle Q_i^{(n)}, K_j\rangle + \omega^{\text{pos}} M^{\text{pos}(n)} - \omega^{\text{neg}} M^{\text{neg}(n)}}{\sum_{k=1}\exp\langle Q_i^{(n)}, K_k\rangle}. \tag{6}$$

It is found that larger weights at higher noise levels [3] can lead to better results, thus ω^{pos} and ω^{neg} are noise-level sensitive parameters, defined by:

$$\omega^{(\cdot)} = \omega' \cdot \log(1+\sigma) \cdot \max(\boldsymbol{A}^{(n)}), \tag{7}$$

where ω' is a user-provided hyper-parameter.

We then substitute the feature map corresponding to the object description. Contrary to the image editing motivation behind conventional prompt tuning methods, it is not necessary in our method to provide a reference image and an amended prompt. Recalling the framework of our generalized ControlNet, the branch architecture is identical to the encoder portion of the backbone network. We found that the attention maps $A^{(n)}_{j|\boldsymbol{\theta}_t^{(n)}}$ within the branch network also encompass object locations and shape information. Consequently, we opt for the attention map $A^{(n)}_{j|\boldsymbol{\theta}_t^{(n)}}$ associated with the description K_j from the respective layer in ControlNet as the source for original attention map $A^{(n)}_{j|\boldsymbol{\theta}^{(n)}}$ substitution, ensuring that the information in the substituted attention map aligns more closely

Table 1: Quantitative comparison on the COCO5k validation set. The best result is highlighted.

Method	Similarity (LPIPS ↓)	Sketch Map (L2 Distance ↓)	Segmentation Map (mPA ↑)	Segmentation Map (mIoU ↑)	Pose Map (mAP ↑)
Multi-Adapter	0.7273 $\pm_{0.00120}$	7.93310 $\pm_{0.01392}$	26.30 $\pm_{0.242}$	13.98 $\pm_{0.177}$	40.02 $\pm_{0.761}$
Multi-ControlNet	0.6653 $\pm_{0.00145}$	7.59721 $\pm_{0.01516}$	36.59 $\pm_{0.273}$	22.70 $\pm_{0.229}$	38.19 $\pm_{0.761}$
Ours w/o ControlNorm	0.4900 $\pm_{0.00141}$	**7.18413** $\pm_{0.01453}$	48.26 $\pm_{0.287}$	32.66 $\pm_{0.272}$	61.93 $\pm_{0.775}$
Ours	**0.4836** $\pm_{\mathbf{0.00133}}$	7.28929 $\pm_{0.01385}$	**49.20** $\pm_{\mathbf{0.289}}$	**33.27** $\pm_{\mathbf{0.271}}$	**61.99** $\pm_{\mathbf{0.778}}$

Figure 4: Our model can generate images with the provided prompts and multi-modality information (e.g., edge, pose, and segmentation map) across various scales.

with the external input signal rather than the textual information derived from the original backbone network. Specifically, we replace the attention map generated by the original backbone network (based on the corresponding object description) with the attention map from the corresponding module in the branch network:

$$\hat{\boldsymbol{A}}^{(n)} = [\tilde{A}^{(n)}_{0|\boldsymbol{\theta}^{(n)}}; \ldots; A^{(n)}_{j|\boldsymbol{\theta}^{(n)}_t}; \ldots; \tilde{A}^{(n)}_{N_t|\boldsymbol{\theta}^{(n)}}] \leftarrow [\tilde{A}^{(n)}_{0|\boldsymbol{\theta}^{(n)}}; \ldots; \tilde{A}^{(n)}_{j|\boldsymbol{\theta}^{(n)}}; \ldots; \tilde{A}^{(n)}_{N_t|\boldsymbol{\theta}^{(n)}}]. \tag{8}$$

Subsequently we can produce the spatially guided output from cross-attention layer by taking the product of $\hat{\boldsymbol{A}}^{(n)}$ with V.

4 Experiments

In this section, we delve into a comprehensive experimental analysis to validate the efficacy and superiority of the proposed method through ablation studies and application demonstrations. Subsequently, in Sec. 4.1, we put forth both quantitative and qualitative results, elucidating the comparative advantages of our approach. We also present an array of intriguing applications made possible by our gControlNet, showcasing its practical utility. Finally, Sec. 4.2 is dedicated to the discussion of ablation studies, scrutinising the impacts of varying injection and sampling methods. The experimental configurations, including the dataset specifications, implementation details, and evaluation metrics, can be found in the Appendix.

4.1 Applications and Comparisions

Cocktail is proficient in seamlessly supporting multiple control inputs and autonomously fusing them, thereby eliminating the necessity for manual intervention to equilibrate diverse modalities. This unique property empowers users to easily incorporate a variety of modalities, resulting in more flexible multi-modal control. In Figure 4, we demonstrate how users can provide spatial information about multiple objects in different modalities to generate a complex scene. Notably, this entire process is accomplished by a single model without the need for additional branch networks.

We then compare Cocktail with two state-of-the-art methods: ControlNet [52] and T2I-Adapter [32], in the context of text-guided image-to-image translation within multiple modalities. We employ several evaluation methods, including LPIPS, mPA, mIoU, mAP and L2 distance for various modalities

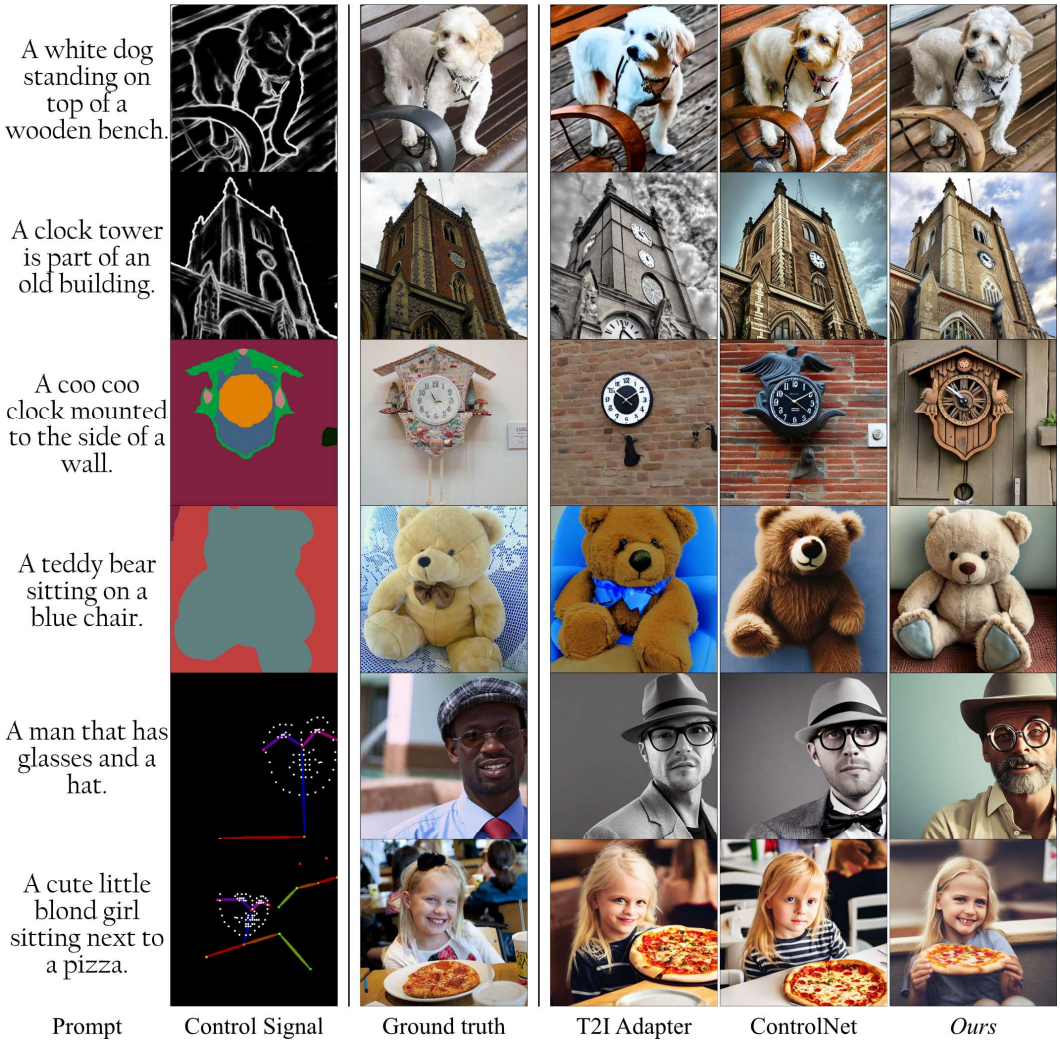

Figure 5: Qualitative comparison of Uni-Modality on the COCO validation set.

in this section. As depicted in Table 1, our method outperformed both ControlNet and T2I-Adapter across all evaluation metrics. This remarkable achievement signifies that our proposed cocktail can generate a structural image that closely resembles the ground truth image and aligns better with the input conditions, establishing its superiority. The visualization in Fig. 6 also illustrates the superior ability of our model to harmonize with the control signals.

We further present some samples from our method to examine its effectiveness on uni-modality translation. The visual comparison on the COCO validation set is showcased in Figure 5, highlighting the compelling performance. Benefiting from mixed training of multiple modalities, our method achieves remarkable generation quality, surpassing even models exclusively trained for a single control signal.

Our experiments show that Cocktail effectively leverages information from different modalities and exhibits outstanding multi-object generation abilities with consistent composition across various control signals.

4.2 Ablations

gControlNet and ControlNorm Previous methods [52] utilized a direct sum approach to fuse control signals from hyper-networks with latent variables from the original network. While this method conveys spatial information, it fails to consider semantic information, such as text or other

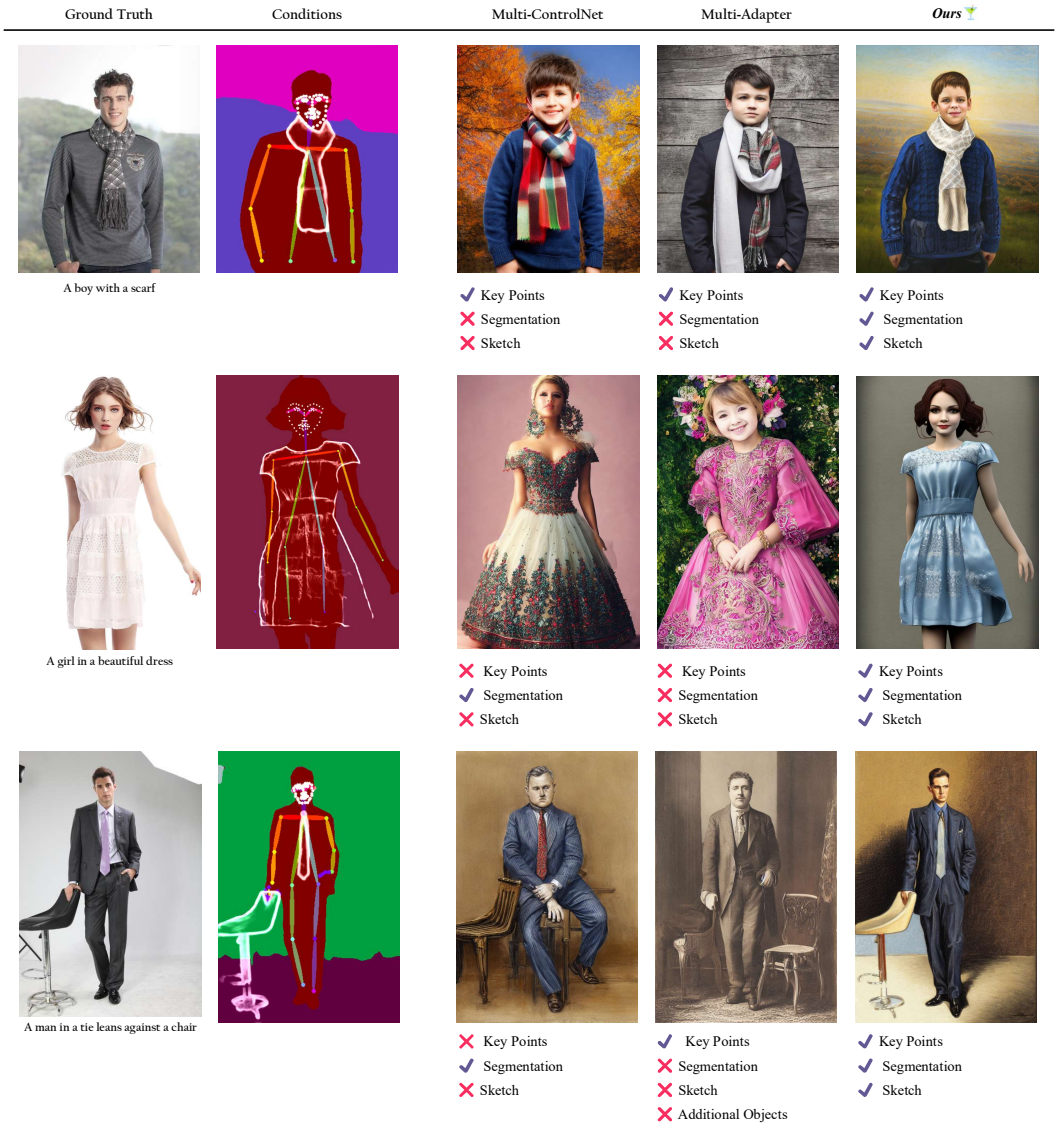

Figure 6: Cocktail can address the imbalance among various modalities.

modalities. Control signals decoupled through ControlNorm allow the preservation of semantic information while conveying spatial information. Another point is the need for normalization to address the imbalance of different modality signals. However, typical normalization methods lead to the loss of semantic information [35]. Therefore, the control signals introduced through ControlNorm can better interpret conditional information. We present some generated images in Fig. 9 to substantiate the interpretative capability of ControlNorm.

Spatial Guidance Sampling The spatial guidance sampling method not only ensures that objects are generated within controllable areas, but also minimizes the impact on other areas. An intuitive application is that when we modify certain objects or modalities, other parts of the generated image can remain unchanged. Figure 7 illustrates the contrast between employing spatial guidance sampling and its absence. A more significant variation in the overall tone of the resultant image is observed when spatial guidance is not utilized, leading to inconsistencies in the details of the attire. Conversely, the incorporation of new objects and modalities with the use of spatial guidance minimally impacts the original image.

5 Conclusion

In this work, we proposed Cocktail, a pipeline to achieve fine control by fusing multi-modal signals. Firstly, we introduced a generalized ControlNet capable of handling signals of different modalities using a single model. We also revealed that the semantic information of signals from the hyper-network may be lost through direct addition. However, a simple decoupling allows for a better interpretation of control signals. We introduced a controllable normalisation for decoupling and integration. Finally, we proposed a sampling scheme that can prevent the generation of unnecessary objects outside the focus area while also achieving a certain degree of editing capability and background protection.

Limitations. While we have achieved multi-modal fusion and control, there are still certain limitations in the handling of control signals from Cocktail that need to be addressed in future work. Firstly, although the current spatial guidance method can effectively prevent objects from being generated outside the focus area, it requires users to individually specify the area and corresponding object description during implementation. Secondly, spatial guidance can cause instability in the latent space in certain situations, leading to the generated images degrading and deviating from the existing control signals. Therefore, finding a stable anchor point as a reference during the generation process is also worth exploring.

Broader Impacts. The integration of multimodal control signals as inputs for synthesis greatly enhances user interaction flexibility and streamlines the utilization of text-conditional diffusion models. However, the process of fine-tuning large-scale generative models to accommodate diverse modalities requires a significant amount of energy, and we anticipate that the development of a universal model capable of accepting multiple modalities will help mitigate this impact. However, the growing capabilities in image generation also facilitate the production of manipulated images with malicious intent, such as the creation of counterfeit or deceitful information.

References

[1] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[2] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, pages 18511–18521, 2022.

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[4] Manuel Brack, Patrick Schramowski, Felix Friedrich, Dominik Hintersdorf, and Kristian Kersting. The stable artist: Steering semantics in diffusion latent space. *arXiv preprint arXiv:2212.06013*, 2022.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2021.

[6] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023.

[7] Ting Chen, Mario Lucic, Neil Houlsby, and Sylvain Gelly. On self modulation for generative adversarial networks. *arXiv preprint arXiv:1810.01365*, 2018.

[8] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[10] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11389–11398, 2022.

[11] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021.

[12] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.

[13] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *arXiv preprint arXiv:2206.09012*, 2022.

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.

[15] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[17] 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.

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[21] Minghui Hu, Chuanxia Zheng, Heliang Zheng, Tat-Jen Cham, Chaoyue Wang, Zuopeng Yang, Dacheng Tao, and Ponnuthurai N Suganthan. Unified discrete diffusion for simultaneous vision-language generation. *arXiv preprint arXiv:2211.14842*, 2022.

[22] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.

[23] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[24] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

[25] Alexander Kirillov. Mscoco keypoint evaluation metric. https://github.com/cocodataset/cocodataset.github.io, 2021.

[26] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[28] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023.

[29] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

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

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

[33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[34] Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-guided diffusion with inside-outside attention. *arXiv preprint arXiv:2212.00210*, 2022.

[35] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[36] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. *arXiv preprint arXiv:2303.11306*, 2023.

[37] 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.

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

[39] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

[40] 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.

[41] 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.

[42] Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. Collage diffusion. *arXiv preprint arXiv:2303.00262*, 2023.

[43] Christoph Schuhmann. Clip+mlp aesthetic score predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

[46] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022.

[47] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.

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

[49] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *International conference on computer vision*, pages 1395–1403, 2015.

[50] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

[51] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2302.12242*, 2023.

[52] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[54] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*, 2022.

[55] Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022.

[56] Jianbin Zheng, Daqing Liu, Chaoyue Wang, Minghui Hu, Zuopeng Yang, Changxing Ding, and Dacheng Tao. Mmot: Mixture-of-modality-tokens transformer for composed multimodal conditional image synthesis. *arXiv preprint arXiv:2305.05992*, 2023.

A Experimental settings

A.1 Datasets and Annotations.

All of our experiments are performed on LAION-AESTHETICS-6.5 dataset, which contains about 600K image-text pairs with predicted aesthetics scores of higher than 6.5.

Our experiment includes three types of commonly used additional control signals, and we use different methods to obtain pseudo-labeling annotations:

- *Sketch map*: The delineation of texture details within the structured control signal is accomplished using the HED boundary detector [49]
- *Semantic segmentation map*: To derive segmentation maps from images, we leverage SAN [51], a high-performing, open-vocabulary segmentator.
- *Human pose map*: The generation of a comprehensive human pose map, including body, face, and hand positions, is facilitated by OpenPose [5].

It is crucial to note that our proposed Generalized ControlNet exhibits broad applicability and can be effortlessly adapted to accommodate a multitude of other modality inputs.

A.2 Implementation Details.

gControlNet is adapted from the pretrained Stable Diffusion v2.1 in this paper and trained for 20 epochs with a batch size of 64 on 4 NVIDIA 80G-A100 GPUs within 4 days. We use the AdamW optimizer with a learning rate of 3.0 e-05. All the training images in the LAION-AESTHETICS-6.5 are first resized to 512 by the short side and then randomly cropped to 512×512. During inference, the sampler is DDIM, the sampling steps are 50, and the classifier-free guidance scale is 9.0 by default.

A.3 Evaluation Metrics.

In order to compare the generation performance of existing methods with ours, we adopted six metrics to evaluate the quality, text-image alignment, aesthetics, and human preference of the generated images. They are Frechet Inception Distance (FID) [17], CLIP/BLIP Score [37, 27], Aesthetic Score [43], Human Preference Score (HPS) [48] and ImageReward [50]. Specifically, FID measures the distribution distances of real and generated image sets utilizing a pre-trained classification network (*e.g.*, Inception V3 trained on ImageNet). Instead, CLIP/BLIP Score calculates cosine similarities between the corresponding image and text features extracted by the CLIP/BLIP image and text encoders, respectively. The Aesthetic Score is based on an additional MLP layer on top of a pretrained CLIP image encoder to measure the human aesthetic aspect of a single image. To consider the alignment with human values and preferences, HPS and ImageReward were proposed and are based on the reward model pretrained on CLIP or BLIP.

However, these metrics do not effectively measure the fidelity of our model to different modalities. Therefore, we additionally employ several other evaluation metrics, including the Learned Perceptual Image Patch Similarity (LPIPS) [53] for overall image quality, the L2 distance for Holistically-Nested Edge Detection (HED), the mean Pixel Accuracy (mPA) [30] and mean Intersection over Union (mIoU) [30] for segmentation maps, and the mean Average Precision (mAP) over 10 object keypoint similarity (OKS) thresholds [25] for human pose map. Specifically, LPIPS is utilized to assess the dissimilarity between the generated image and the ground-truth image. Furthermore, we extract the conditions, namely the sketch map, segmentation map, and pose map, from both the generated image and the ground-truth image. By computing the distance, specifically employing L2 Distance, mPA, mIoU, and mAP, between these extracted conditions, we can gain a deeper understanding of the fidelity with respect to various modalities.

Through these metrics, we can more clearly ascertain whether the generated images can follow the guidance of different modal conditions.

B Additional quantitative analysis

We compare Cocktail with two state-of-the-art methods: ControlNet [52] and T2I-Adapter [32], in the context of text-guided image-to-image translation within a single modality. As demonstrated in Table 2, our model not only acquires multi-modal capabilities but also demonstrates performance on par with the comparative methods in single-modal tasks. It is worth noting that FID was measured on the basis of 5000 images from zero-shot generation.

It worth noting that FID may not provide a precise evaluation as they are substantially contingent on the sample size and the characteristics of the training dataset. Our methodology prioritizes the fidelity of the generated images, a perspective that is at odds with the principles of CLIP/BLIP.

C Additional qualitative results

We provide additional generated images under various multi-conditional scenarios and compare them with Multi-ControlNet and Multi-Adapter. In Fig. 6, we have demonstrated some samples within less complexity, we further show the comparative performance of our model against other state-of-the-art models under challenging disjoint scenarios in Fig. 8. In Fig. 9, we also conduct an ablation study on the effectiveness of our ControlNorm module. It can be observed that our method outperforms other state-of-the-art models in terms of balancing and expressing multi-modal signals, and ControlNorm provide a more stable integration performance.

Table 2: Additional quantitative comparison on the COCO validation set. The best result is highlighted.

(a) Text + Edge

Method	FID ↓	CLIP Score ↑	BLIP Score ↑	Aesthetic Score ↑	HPS ↑	ImageReward ↑
ControlNet	18.73	0.2588	**0.5338**	**5.39**	19.53	**0.3726**
T2I-Adapter	20.21	**0.2638**	0.5335	5.36	**19.69**	0.3344
Ours	**16.66**	0.2561	0.5321	5.35	19.53	0.2989

(b) Text + Segmentation Map

Method	FID ↓	CLIP Score ↑	BLIP Score ↑	Aesthetic Score ↑	HPS ↑	ImageReward ↑
ControlNet	**20.53**	**0.2640**	**0.5262**	5.43	19.85	0.2609
T2I-Adapter	21.81	0.2626	0.5225	5.25	19.57	0.0837
Ours	23.88	0.2613	0.5260	**5.64**	**19.89**	**0.3735**

(c) Text + Pose

Method	FID ↓	CLIP Score ↑	BLIP Score ↑	Aesthetic Score ↑	HPS ↑	ImageReward ↑
ControlNet	**35.13**	0.2656	0.5171	5.49	20.17	0.3683
T2I-Adapter	35.53	**0.2697**	**0.5211**	5.45	20.19	0.3810
Ours	39.64	0.2641	0.5210	**5.70**	**20.22**	**0.4488**

(d) Text + Sketch + Segmentation + Keypoints

Method	FID ↓	CLIP Score ↑	BLIP Score ↑	Aesthetic Score ↑	HPS ↑	ImageReward ↑
Ours	16.70	0.2562	0.5326	5.36	19.56	0.3118

Figure 7: With spatial guidance, our method is capable of maintaining constancy in certain regions while modifying specific objects.

We also provide further samples under single modality signals, *e.g.*, Human Pose in Figs. 10&11, Sketch Map as shown in Figs. 12&13 and Segmentation Map in Fig. 14. It can be seen that our method is more faithful to the given conditional signals.

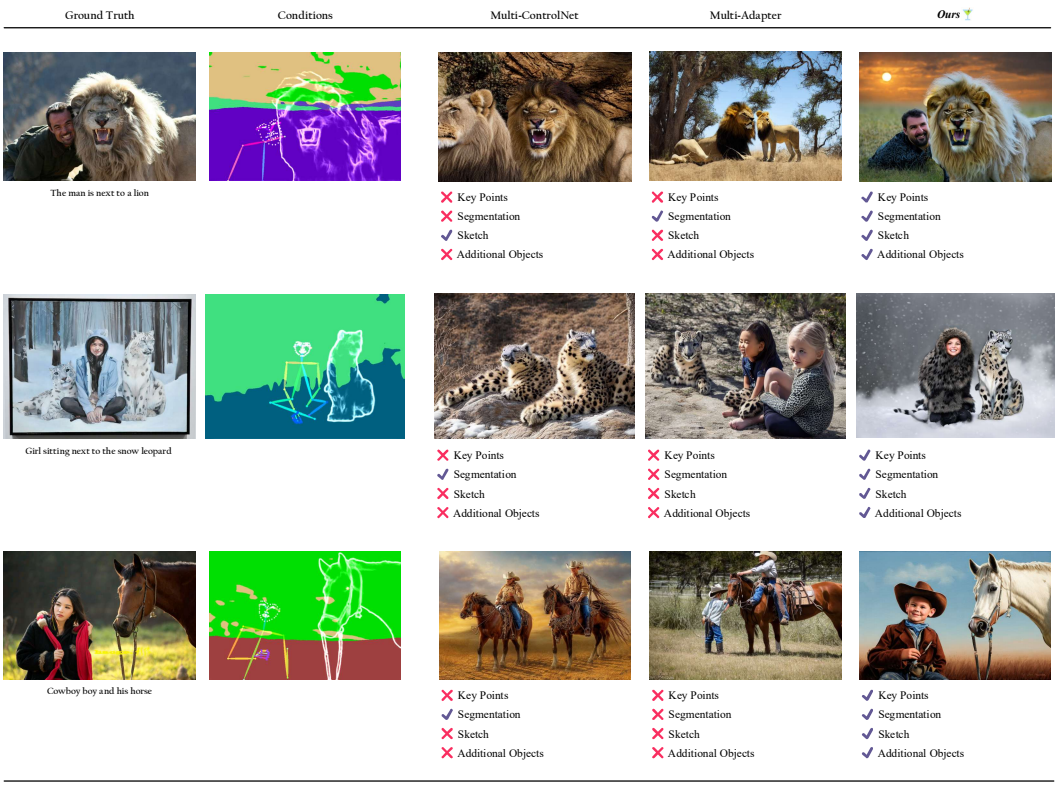

Figure 8: Additional results for Disjoint Multi-Modality Control.

Figure 9: ControlNorm can address the imbalance among various modalities. Note that the framed areas in magenta and cyan do not provide a high-fidelity interpretation.

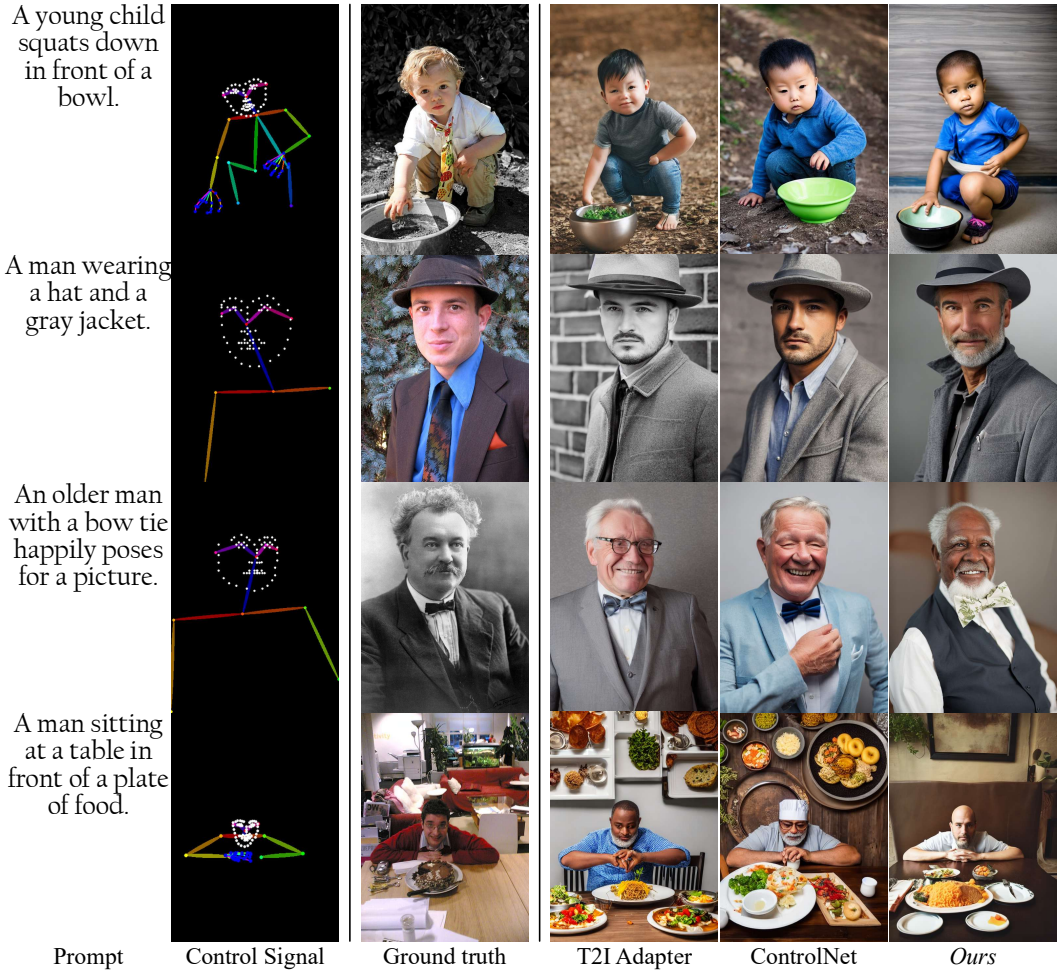

Figure 10: Additional results for Single Modality Control: Key Points

Figure 11: Additional results for Single Modality Control: Key Points

Figure 12: Additional results for Single Modality Control: Sketch

Figure 13: Additional results for Single Modality Control: Sketch

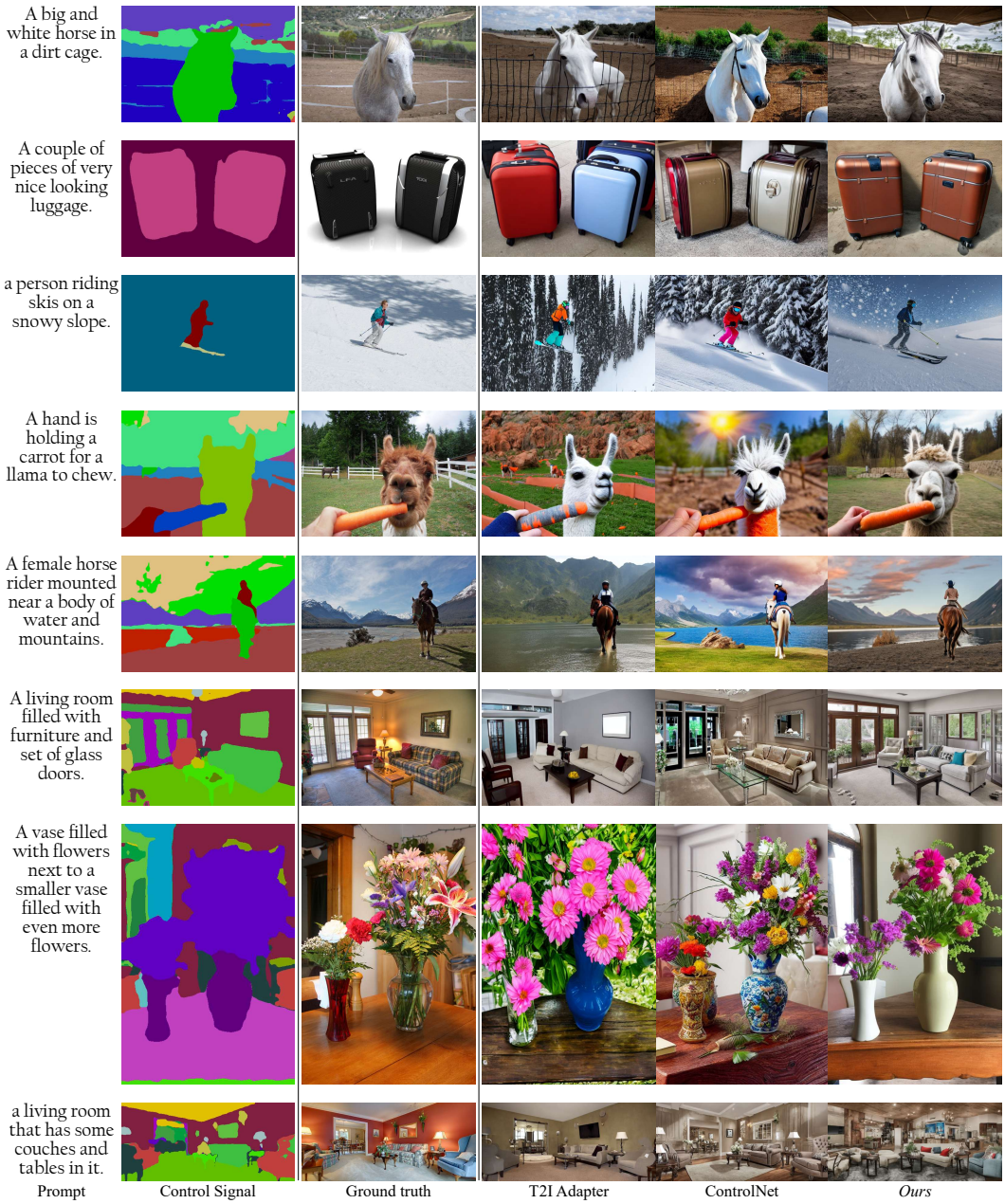

Figure 14: Additional results for Single Modality Control: Segmentation Map