Text-Anchored Score Composition: Tackling Condition Misalignment in Text-to-Image Diffusion Models

Luozhou Wang[1⋆], Guibao Shen[1⋆], Wenhang Ge[1], Guangyong Chen[3,4], Yijun Li[5], and Yingcong Chen[1,2⋆⋆]

[1] Hong Kong University of Science and Technology (Guangzhou)
[2] Hong Kong University of Science and Technology
[3] ZhejiangLab
[4] Zhejiang University
[5] Adobe Research

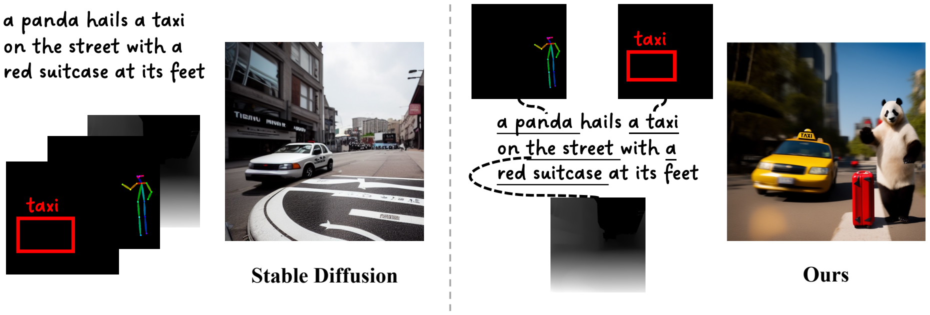

a panda hails a taxi on the street with a red suitcase at its feet
taxi
Stable Diffusion
taxi
a panda hails a taxi on the street with a red suitcase at its feet
Ours

Fig. 1: Illustration of our proposed **Text-Anchored Score Composition** showcasing the ability to handle the misalignment between conditions for controllable generation tasks. For example, with several control conditions (e.g., depth, pose, bounding box) specifying the layout structure and the text condition indicating **extra** guidance (e.g., new object, new correspondence of spatial relationship), our method is able to generate high-quality plausible outputs that satisfy all given conditions without training.

Abstract. Text-to-image diffusion models have advanced towards more controllable generation via supporting various additional conditions (e.g., depth map, bounding box) beyond text. However, these models are learned based on the premise of perfect alignment between the text and extra conditions. If this alignment is not satisfied, the final output could be either dominated by one condition, or ambiguity may arise, failing to meet user expectations. To address this issue, we present a training-free approach called **Text-Anchored Score Composition (TASC)**

⋆ Equal contribution
⋆⋆ Corresponding author

to further improve the controllability of existing models when provided with partially aligned conditions. The TASC firstly separates conditions based on pair relationships, computing the result individually for each pair. This ensures that each pair no longer has conflicting conditions. Then we propose an attention realignment operation to realign these independently calculated results via a cross-attention mechanism to avoid new conflicts when combining them back. Both qualitative and quantitative results demonstrate the effectiveness of our approach in handling unaligned conditions, which performs favorably against recent methods and more importantly adds flexibility to the controllable image generation process.

Keywords: Controllable Image Generation· Condition Misalignment

1 Introduction

Diffusion models [8,14,27,27,30,30,31,33,33–35], epitomized by Stable Diffusion (SD) [31], are notably proficient in image synthesis. Central to their functioning, and a principle they share with score-based models [36, 37], is the prediction of the underlying score of the data distribution. Building upon this, advancements have been made by extending SD with additional conditions, specifically through the implementation of ControlNet [45], the T2I Adapter [26] and GLIGEN [20]. These innovations enhance controllability with the idea of the adapter, thereby obviating the need for training from scratch.

However, the increased complexity of control signals originates another challenge: multi-condition controllable image generation relies on the assumption that all conditions are well-aligned. This strict assumption limits the scenarios in which users can effectively employ controllable generation. The misalignment of multiple control conditions can negatively impact the overall performance and user experience during controllable image generation. This misalignment typically results in two phenomena: (i) *Dominance*, where one condition takes over the output; (ii) *Ambiguity*, leading to unclear or conflicting outcomes. For instance, consider the case where the text condition mentions two objects, but the extra condition contains only one object. Dominance occurs when the extra condition accurately corresponds to one of the objects in the text, leading the generation process to be dominated by the extra condition and neglecting the other object. As illustrated on the left side of Figure 2, the dog in the text "a car and a dog" is omitted. This phenomenon happens across different control methods [20, 26, 45]. On the other hand, ambiguity arises when the extra condition could refer to either one of the objects in the text, resulting in an unclear correspondence. In this situation, it becomes challenging to successfully generate an image where the extra condition accurately corresponds to a specific object mentioned in the text, as designated by the user. As depicted on the right side of Figure 2, The user is unable to control which object to generate simply based on the depth map.

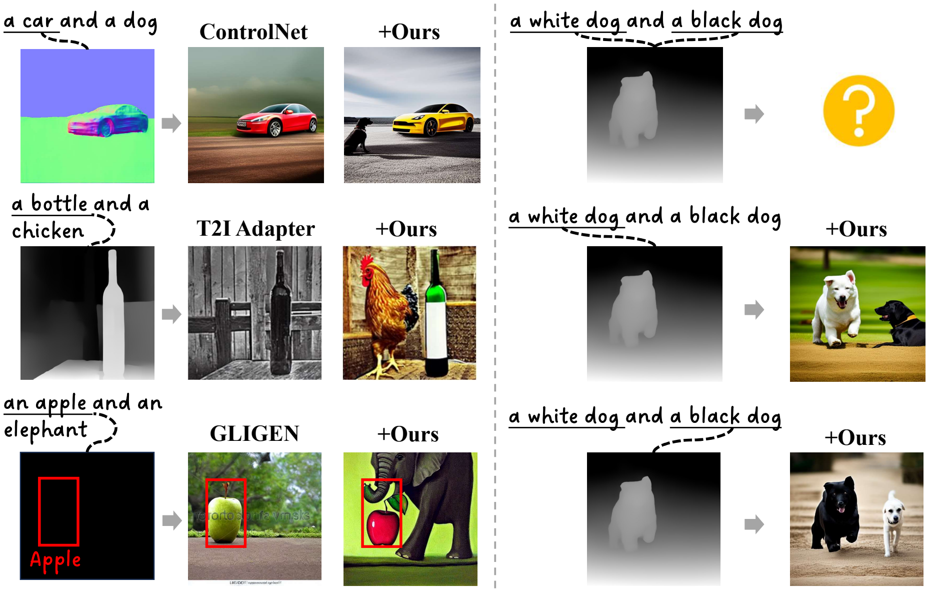


Fig. 2: Challenges in Multi-Condition Image Synthesis. Left: the "dominance" effect exists across various methods, where one object (*e.g.*, car) overshadows the generation, omitting the other (*e.g.*, dog). Right: The "ambiguity" issue, where it's unclear which object from the text matches the intended depth condition.

In this work, we introduce a training-free approach to address the misalignment between text and image conditions for a broader range of controllable image generation needs. We propose the **Text-Anchored Score Composition (TASC)**, which treats the entire text as a unified condition, with each additional condition corresponding to a different subset of words in the text. Firstly, our score composition formula separates conditions based on pair relationships by leveraging the fine-grained needs of users. During each timestep of the diffusion generation process, rather than inputting all conditions together into the denoising network, we calculate the model output for each condition pair separately, termed the "individual score." Conversely, the "unified score" is computed using the entire text input as a single condition. This process ensures that the impact of each condition remains unaffected by the dominance of other conditions. Nevertheless, the computation process of these scores can be further optimized. We observed that at each step of diffusion, the calculations for individual scores and the unified score are independent, potentially leading to conflicts between the unified and individual scores. To address this, we propose an attention alignment operation, aimed at resolving the conflict between the unified and individual scores. This operation realigns the attention values of the unified score with those of the individual scores, resulting in a modified unified score. This modified unified score is then combined with the individual scores and used for diffusion sampling to produce the final output.

Our Text-Anchored Score Composition (TASC) offers a versatile and comprehensive solution for controllable image generation, addressing various needs while reducing potential conflicts and overlaps between conditions as shown in Figure 1. In conclusion, our contributions can be summarized as follows:

- We aim at enabling more flexible controllable generation when provided with partially aligned conditions, which is significantly different from the conventional need for perfectly aligned conditions in previous works.
- We propose a training-free approach and demonstrate its effectiveness with extensive experiments to efficiently compute and align both individual and unified score estimates, generating high-quality results that meet user expectations.

2 Related Work

Text-to-Image Diffusion. Diffusion model has emerged as a powerful technique for generating high-quality images [27,30,31,33,43]. The Stable Diffusion (SD) [31] model, which leverages latent diffusion processes, serves as a prime example. This text-to-image diffusion model becomes the foundation for generation of other modality and tasks [21,29,39,40]. Further, certain studies [10,16,17,32] have expanded on this foundation, incorporating image guidance into the diffusion process for customization. By binding specific image features with the text, these models simulate the corresponding style through the text prompt. Others [3,7,24,44,46] focus on exploiting spatial information inherent in images, integrating the image directly into the computations to ensure better alignment in the final output.

Multi-Condition Control. The ability to integrate multiple control conditions into existing models has been a notable advancement, facilitated by techniques like ControlNet [45] or T2I adapter [26]. These methods enable diffusion models like Stable Diffusion to handle multiple image conditions without the need for extensive retraining. These methods support a variety of condition types, such as depth, normal, human pose, and canny, thereby expanding the scope of image generation tasks. However, the introduction of multiple conditions leads to the issue of condition composition. **Although previous works, like [2,4,23,41], have addressed the composition of multiple conditions during generations, none of them have discussed the potential misalignment among these conditions.** Building upon these approaches, our work is the first attempt to address this problem, providing a training-free approach to unify the misaligned conditions, and thus significantly improve their controllability.

Cross-Attention Modifications. The cross-attention mechanism has become a cornerstone of the text-to-image diffusion model, with several studies [12,25,28,38] employing this mechanism to execute a variety of image editing tasks. Notably, altering attention values has been suggested as an effective strategy for steering image generation, as exemplified in models such as StructureDiffusion [9] and Attend-and-Excite [5]. During the generation process, Attend-and-Excite [5] modifies the latent at each timestep by maximizing the attention value of designated tokens, thereby ensuring the synthesis of both objects in the final output. Conversely, StructureDiffusion [9] disassembles the text input into hierarchical levels and consolidates them through cross-attention computation. Several studies have also employed cross-attention mechanisms to precisely regulate the positioning of specific contents or objects within images [6, 11, 18, 42]. Nevertheless, these methodologies exhibit diminished efficacy in addressing the dominance issue when confronted with multiple conditions. More critically, they often stumble upon ambiguity, as illustrated in Figure 2, where users are incapable of controlling the specific region for object generation.

3 Proposed Method

3.1 Preliminary

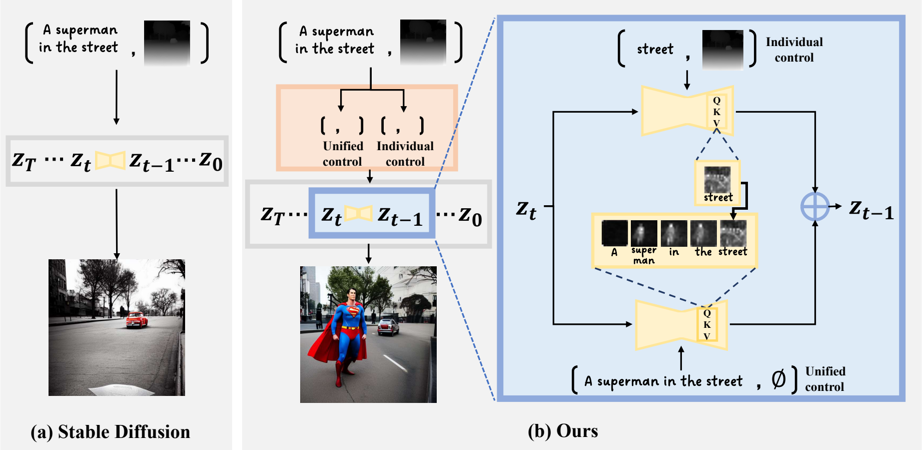
Multiple-Conditioned Image Generation. Our multi-condition framework requires additional parameters, such as distinct sets ϕ for different conditions (e.g., ControlNet [45] or GLIGEN [20]). When computing scores, it's crucial to specify the active conditions, even if some are inactive. To streamline notation for active conditions, we use simplified expressions. For instance, $\epsilon_{\theta,\phi_2}(z_t, \varnothing, \mathcal{I})$ represents a scenario with only the additional condition $\mathcal{I}$ active, maintaining the text condition slot in our notation regardless of use. This approach helps distinguish our method from others like Composable Diffusion [23], which doesn't fully consider the complex interactions between conditions and text in multi-condition settings.

3.2 Condition Misalignment

Given a text prompt $\mathcal{P}$ and additional conditions $\mathbb{I} = [\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_K]$, the score of the original classifier-free guidance [15] is computed as follows:

$$\begin{aligned}\epsilon_{\text{CFG}}(z_t, \mathcal{P}, \mathbb{I}) &= \epsilon_\theta(z_t, \varnothing, \varnothing) \\ &+ w \cdot (\epsilon_{\theta,\phi_{1,\dots,K}}(z_t, \mathcal{P}, \mathbb{I}) - \epsilon_\theta(z_t, \varnothing, \varnothing)).\end{aligned} \tag{1}$$

Often times we feed conditions into the model collectively, without addressing potential misalignment issues, *i.e.*, the dominance and ambiguity. We introduce and analyze these two phenomena through the example with $K = 1$ as shown in Figure 2. **Dominance** represents a form of partial alignment, where additional condition $\mathcal{I}_1$ corresponds solely to one object $\mathcal{P}_i$ in the overall text prompt $\mathcal{P}$. However, when these conditions are simultaneously fed into the model, it struggles to concurrently fulfill the text and additional conditions. Consequently, the output exhibits a dominance phenomenon, with one object receiving precedence while the others are disregarded. **Ambiguity** emerges as a by-product of dominance, wherein the additional condition might correspond to any object within the text, *i.e.*, $\mathcal{I}_1$ could either correspond to $\mathcal{P}_i$ or $\mathcal{P}_j$. Thus the conditions fed into the model create a potential for ambiguity. Moreover, in practical applications, even when users delineate the correspondence, current methods find it challenging to leverage this information effectively.


A superman in the street ,
$z_T \cdots z_t$ $z_{t-1} \cdots z_0$
Unified control
Individual control
street ,
Q K V
street
A super man in the street
z_t
z_{t-1}
A superman in the street , ∅

(a) Stable Diffusion

(b) Ours

Fig. 3: Image Synthesis using Text-Anchored Score Composition. Given a depth map of a street and a textual description specifying "a Superman in the street", our method effectively generates the corresponding image.

3.3 Our Solution

To address the problem discussed earlier, we propose a Text-Anchored Score Composition. Our solution consists of two components: the *Text-Anchored Score Composition formula* , which breaks down the conditions into fully-aligned pairs, and the *attention realignment*, which ensures that the decomposed results are more consistent when combined together.

Text-Anchored Score Composition. Initially, we assume users will specify pair relationships $S(\cdot)$ to setting the details of individual entities in the text, using image conditions or bounding boxes, like setting a dog's color as shown in Figure 2. We presume users won't provide meaningless inputs, ensuring pair relationships are practical and aligned with their creative goals.

This function formalizes the pairing between additional conditions and text tokens. By this function, we can establish aligned condition pairs $\{(\mathcal{P}_{S(k)}, \mathcal{I}_k)\}_{k=1}^{K}$. Leveraging these pairs, we introduce the **Text-Anchored Score Composition** (TASC) equation:

$$
\begin{aligned}
\epsilon(z_t, \mathcal{P}, \mathbb{I}) &= \epsilon_\theta(z_t, \varnothing, \varnothing) \\
&+ w_0 \cdot \underbrace{(\epsilon_\theta(z_t, \mathcal{P}, \varnothing) - \epsilon_\theta(z_t, \varnothing, \varnothing))}_{\text{unified control}} \\
&+ \sum_{k=1}^{K} w_k \underbrace{(\epsilon_{\theta,\phi_k}(z_t, \mathcal{P}_{S(k)}, \mathcal{I}_k) - \epsilon_\theta(z_t, \varnothing, \varnothing))}_{\text{individual control}}.
\end{aligned} \tag{2}
$$

We ensure the preservation of the original text $\mathcal{P}$ as the sole input, serving as a unified control. This strategy ensures the successful generation of all objects,

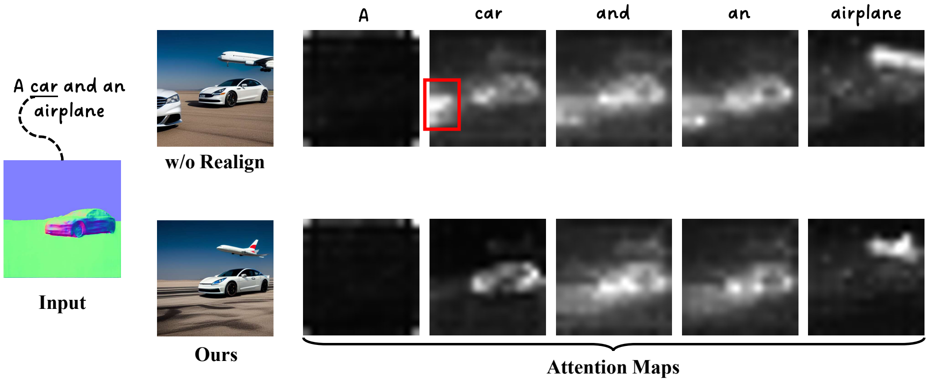


Fig. 4: Illustration of the "Attention Realignment" operation. The operation addresses mismatches between the unified score and individual scores during generation, preventing the final object from appearing multiple times, as individual score for the same token "car" will not contribute to the attention value in the red box.

even those without corresponding additional conditions. Note if without these pair relationships $S(\cdot)$, our method just defaults to standard image-conditioned generation.

Through the equation, we generate individual scores for each pair, eliminating any existing misalignment issues within the computation of individual scores. Both text and additional conditions, therefore, contribute to the final image. Additionally, the implementation of the pair relationship function $S(\cdot)$ allows us to further mitigate any ambiguity based on the specific needs of the user.

However, this equation introduces a potential problem: the unified score and individual scores are calculated independently, which may lead to discrepancies between them. Consequently, the synthesized image might contain some undesirable content, as depicted in Figure 4. By analyzing the attention map during the generation process, we find that the attention map for the "car", which is already constrained by the depth map condition, also has high values in other areas. Intuitively, when generating using individual scores and unified scores separately, the value corresponding to the "car" token on the attention map in the individual score always aligns with the depth map condition, thus restricting it to the correct area. Conversely, the "car" in the unified score is unconstrained. We therefore deduce that this phenomenon is due to the lack of consensus between the unified score and the individual scores on the same token.

Attention Realignment. To ensure appropriate alignment between the unified score and the corresponding individual scores, we introduce an attention realignment operation. During the computation of the individual and unified scores, we can obtain the corresponding attention maps M^0 and $\{M^k\}_{k=1}^K$. We then replace the attention values of the corresponding tokens $S(k)$ with the attention values derived from the k-th individual score. Specifically, $M^0_{S(k)} \leftarrow M^k$ for $k = 1, \ldots, K$, where $M^0_{S(k)}$ is the attention map of the $S(k)$-th token in

$\epsilon_\theta(z_t, \mathcal{P}, \varnothing)$, and M^k is that of $\epsilon_{\theta,\phi_k}(z_t, \mathcal{P}_{S(k)}, \mathcal{I}_k)$. This leads to aligned unified control.

We then substitute this $\epsilon_\theta(z_t, \mathcal{P}, \varnothing)$ into the unified control of Eq. (2) to obtain our final score $\tilde{\epsilon}$, which is subsequently used for sampling. The modified unified score can prevent the generation of undesirable content during the generation process. Additionally, the motivation behind replacing attention is to ensure appropriate interactions; that is, once the car's information is confirmed, the generation of the airplane must be reasonable and should not occupy the car's attention.

Detailed descriptions and mathematical formulations of these operations are provided in the supplementary material. The entire procedure is defined in Algorithm 1.

Algorithm 1 Text-Anchored Score Composition.

```
Input: unified text prompt $\mathcal{P}$, additional conditions $\mathbb{I}$, pair relationship function
  $S(\cdot)$, pretrained diffusion model with $K$ controller model parameters $\epsilon_{\theta,\phi_{1,\dots,K}}$
Initialize sample $z_T \sim N(0, 1)$
for $t = T$ to 1 do
    $\epsilon, \_ \leftarrow \epsilon_\theta(z_t, \varnothing, \varnothing)$
    for $k = 1$ to $K$ do
        $\epsilon^k, M^k \leftarrow \epsilon_{\theta,\phi_k}(z_t, \mathcal{P}_{S(k)}, \mathcal{I}_k)$
    end for
    $\epsilon^0, \_ \leftarrow \epsilon_\theta(z_t, \mathcal{P}, \varnothing)\{M^k\}_{k=1}^K$        ▷ attention realignment using $\{M^k\}_{k=1}^K$
    $\tilde{\epsilon} = \epsilon + w_0 \cdot (\epsilon^0 - \epsilon) + \sum_{k=1}^K w_k \cdot (\epsilon^k - \epsilon)$
    $z_{t-1} \leftarrow \tilde{\epsilon}, z_t$        ▷ sampling
end for
```

Discussions. Our method effectively processes complex text prompts and a variety of image control signals, as illustrated in Fig. 1. When calculating a unified score for lengthy text prompts, the issue of missing objects may still arise. Nonetheless, our approach is fully compatible with existing solutions for missing objects, such as Attend & Excite [5], demonstrating its adaptability and ensuring comprehensive object generation.

Furthermore, we recognize that condition conflicts can arise with various adapter combinations, including but not limited to ControlNet [45], GLIGEN [20], and T2I Adapters [26]. Our novel score composition formula, presented in Equation (2), allows us to effectively address these conflicts, as demonstrated in Fig. 5.

4 Experimental Results

4.1 Evaluation Setup

Our experiments employed widely-used controllable methods, specifically those in [20, 26, 45]. Given the similar functionalities of the T2I adapter [26] and ControlNets [45], we chose ControlNets for our study. The image conditions primarily

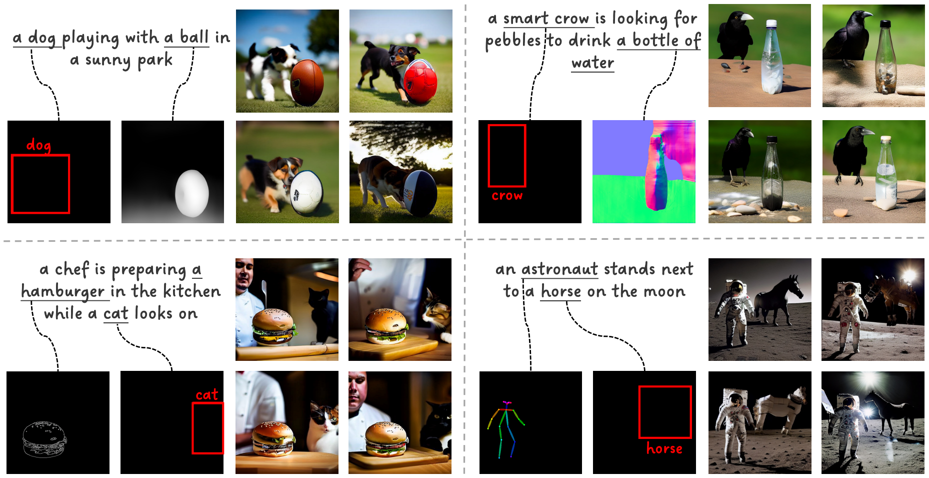

a dog playing with a ball in a sunny park
dog
a smart crow is looking for pebbles to drink a bottle of water
crow
a chef is preparing a hamburger in the kitchen while a cat looks on
cat
an astronaut stands next to a horse on the moon
horse

Fig. 5: Applications. Demonstrating scalability, our method adeptly manages complex scenes, seamlessly combining controller mechanisms like both ControlNet [45] and GLIGEN [20] simultaneously.

included depth, normal, canny, and pose. Additionally, we incorporated grounding token-controlled image generation as described in [20]. The evaluation set, following [5], comprised prompts like "a $[objectA]$ and a $[objectB]$." Each prompt included a control condition for one object, with the other generated textually. Moreover, testing scenarios include specialized prompts for "pose" evaluations, such as **"a [character] and an [object]" and "a [character A] and a [character B],"** with each character depicted in a variety of poses. In total, 444 and 448 unique prompts were used for image conditions and grounding tokens, generating 10 images each. Consider the diversity of prompts, we have significantly expanded the **diversity to include various control types and control images**, in addition to textual prompts.

4.2 Qualitative Evaluation

Applications. As seen in Figure 5, our method proves effective in complex scenes with various conditions. Take the intricate prompt "a smart crow is looking for pebbles to drink a bottle of water," involving three entities (crow, pebbles, bottle) and control signals for two (crow, bottle). Our method adeptly generates all three entities. Importantly, in multi-condition scenarios, conflicts can emerge both between extra conditions (like control image and grounding token) and the text, as well as among the conditions themselves, heightening the complexity of these scenes.

Comparisons of "dominance" Challenge. The results depicted in Figure 6 illustrate our method's effectiveness. It reliably generates every object described in the text across diverse controllable methods, surpassing baseline techniques.

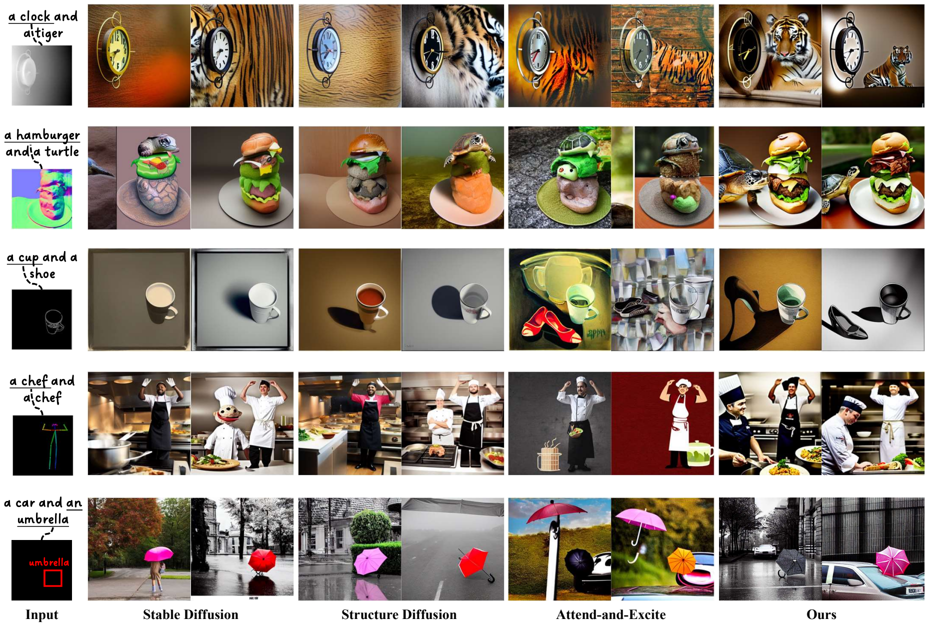


Fig. 6: Comparisons of "Dominance" Challenge. Our approach ensures that objects in the text, not linked to additional conditions, are still accurately generated (*e.g.*, "tiger" in the first example).

Moreover, it maintains a coherent and harmonious composition within each scene.

Comparisons of "ambiguity" Challenge. Our method effectively binds condition to one of the text elements, thereby eliminating this phenomenon. While for baseline methods, which lack this capability, we rewrite the text to reduce ambiguity. For example, given an image with a pose on the right, we add directional words to the text, like "a man on the left and a robot on the right." However, even with text modification, other methods struggle to accurately match the specifics, whereas our method correctly assigns the specified poses to the characters in the text, as shown in Figure 7.

4.3 Quantitative Evaluation

Image-text similarity. For textual alignment, we compute the cosine similarity between the input text prompt and the generated image, which is called the CLIP Score. Furthermore, We pivot to generate matching image captions using a pre-trained BLIP [19] and then compute the average CLIP similarity between the original prompt and all generated captions, which is called BLIP score. Both of the metrics are the higher the better. Table 1 presents our quantitative results. We demonstrate superior performance across different controllable methods. For

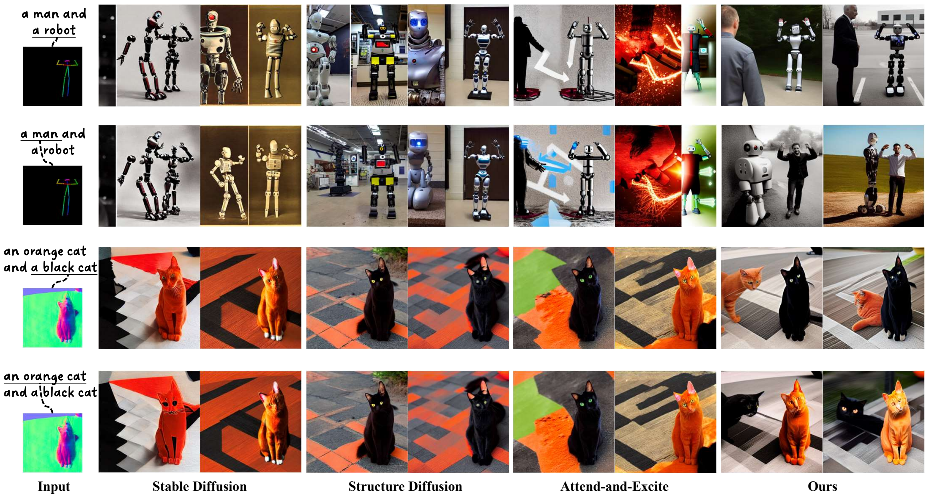


Fig. 7: Comparisons of "Ambiguity" Challenge. Our method accurately maps specific words in the text (*e.g.*, "orange cat" in the last row) to the corresponding additional conditions ("normal map").

Table 1: Quantitative Comparison Results.

Method	CLIP Score	BLIP Score	FID	Time (Seconds)	Memory (MB)
SD	0.2934	0.6949	112.89	**9.70**	**7260.0**
SSD	0.2928	0.6940	112.57	10.87	8730.0
A&E	0.3253	0.7334	111.41	19.36	18840.2
Ours	**0.3323**	**0.7388**	**107.39**	13.88	9764.0

StructureDiffusion [9], we observe a marginal decrease in metrics compared to Stable Diffusion, an observation congruent with the findings in [5].

Image Quality and Diversity. For image Quality and diversity, we adopt the widely used Fréchet Inception Distance(FID) [13] tested on MSCOCO dataset [22] for evaluation. It computes the distance between the distribution of the synthetic images and that of the real test images dataset. The FID is better when lower. Table 1 demonstrates that our proposed method outperforms other baseline methods by a significant margin in terms of image quality.

Computation Cost. We also report both GPU memory usage and inference time of each method in Table 1. Note that all the approaches bring additional costs to get better generation results inevitably compared with Stable Diffusion. Thus our costs are acceptable when considering the significant improvement we contribute.

Impact on Controllability with Different Methods. In our experiments, both our method and A&E managed to generate additional objects, but sometimes struggled with control precision, as seen with A&E's canny controls in

Table 2: Impact on Controllability with Different Methods. Low is better.

	with ControlNets [45]	with GLIGEN [20]
A&E [5]	14.1%	7.36%
Ours	**1.6%**	**4.34%**

Table 3: User study results.

	SD [31]	SSD [9]	AE [5]	Ours
User preference	11%	7%	11%	**71%**

Figure 6. We followed controllability evaluations from [47] and assessed control effectiveness by calculating L2 distances from generated image conditions, and CLIP scores for object placement with GLIGEN. Despite occasional challenges, our method showed lower relative image-condition distances in Table 2, indicating better control retention while adding objects.

4.4 User Study

To rigorously assess the effectiveness of our approach, we organized a comprehensive user study. Participants, numbering **81** in total, were shown 6 prompts randomly paired with various control conditions. This setup generated **10** unique scenarios where condition misalignment was a notable challenge. In each case, participants were presented with a set of 4 images synthesized by different methods under the same condition combinations. The core question of the survey was to identify which set of images most accurately reflected the intended combination of text and control image conditions. Remarkably, 71% of the participants found that the images generated by our method best adhered to the specified conditions, as detailed in Table 3. This overwhelming preference highlights our method's superior capability in effectively addressing and rectifying condition misalignments, thereby validating its practical applicability and user satisfaction in real-world scenarios.

Table 4: Ablation Study.

	with ControlNets [45]		with GLIGEN [20]	
	Simialrity↑	Distance↓	Simialrity↑	Distance↓
SD	69.49%	-	71.24%	-
Ours-w/o-realign	73.78%	9.59%	74.81%	6.23%
Ours	**73.88%**	**1.58%**	**79.87%**	**4.34%**

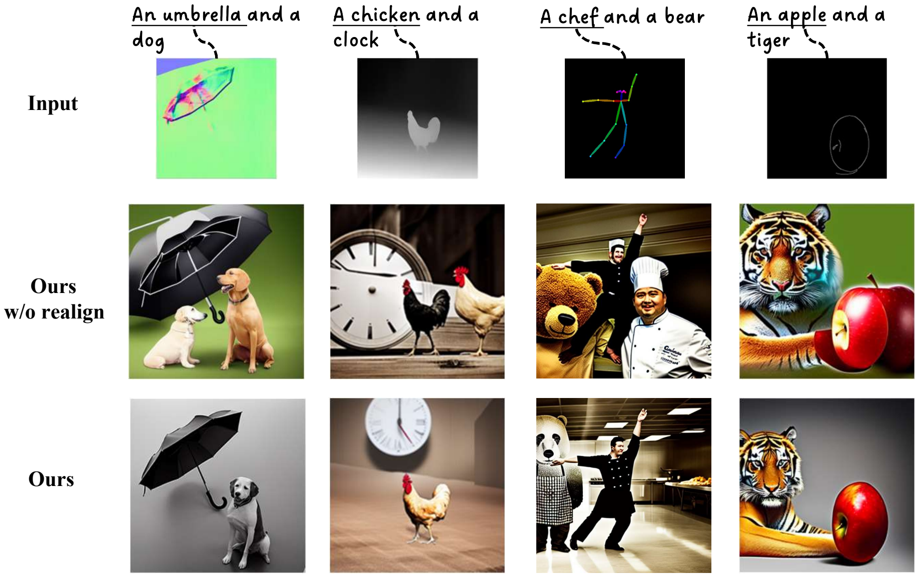


Fig. 8: Ablation Study. Incorporating attention realignment operation into our method prevents the generation of extraneous objects, as exemplified by the absence of redundant elements like the second chef in the column of "pose".

4.5 Ablation Study

We carried out ablation studies using Stable Diffusion with ControlNets or GLIGEN. The results, presented in Table 4 and Figure 8, reveal that our TASC significantly boosts image-text similarity, effectively generating an additional object amidst dominance. Note that the "ours-w/o-realign" method can be considered an enhanced version of Composable Diffusion [23], effectively eliminating any condition misalignment issues in all scoring computations. However, excluding attention realignment operation led to issues in object generation under image conditions, as shown in Figure 8, and resulted in poorer performance in relative image-condition distance. Our full method excels in balancing relative image-condition distances and image-text similarity, producing images that adeptly meet both text and image condition constraints, demonstrating a harmonized approach.

4.6 Discussions

Our method unifies multiple model consensuses through the cross-attention mechanism, ensuring that objects in the text are generated correctly according to model priors. This contrasts with stepwise generation approaches, as illustrated in Figure 9, which require manual intervention (like providing masks) and may struggle in complex scenarios (e.g., "bear behind bottle"). Our approach

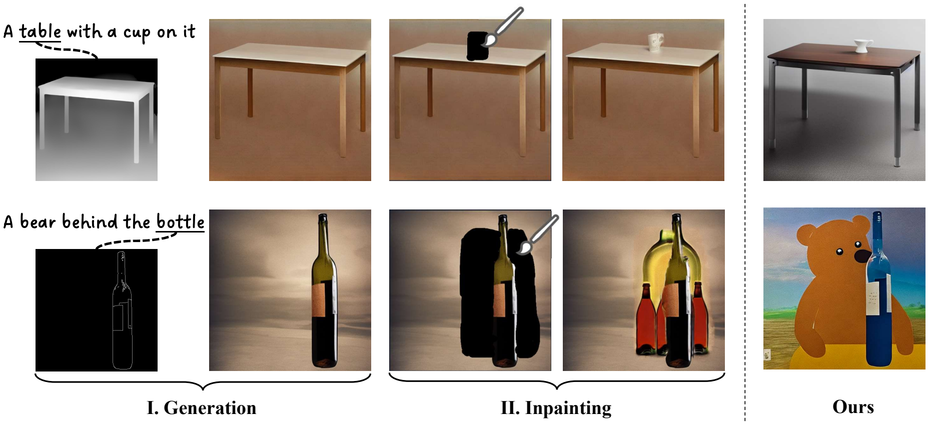


Fig. 9: Left: Two-stage method. Right: Our Approach. Our method autonomously positions objects from the text in their correct locations, leveraging the generative model's inherent priors, without the need for manual intervention such as masking. Inpainting is performed using [1].

effectively leverages model priors. It is noted that the effectiveness of our "Realign" strategy is limited by the model's cross-attention control capability. If the model has a bias in understanding certain words (where the attention map does not fully correspond to the correct text), the alignment efficacy using words as bridges is impacted.

5 Conclusion and Limitation

In summary, our method significantly improves the control capabilities of text-to-image diffusion models. It adeptly manages unaligned conditions, outperforming recent methods in efficiency and effectiveness. This training-free, two-phase approach allows for more intricate and flexible image synthesis, advancing the domain of controllable image generation. Future work could involve identifying a more effective unified control signal. While we primarily utilized text in this paper, scene graphs, with their superior semantic structure, might offer improved control. Their current generation efficacy is limited due to dataset size constraints. As for limitations, our method brings additional computation costs like other state-of-the-art methods and may limit the practical usage to a certain degree.

Acknowledgments

We acknowledge the contributions of the Zhejiang Lab for providing essential GPU resources. Finally, we extend our gratitude to the anonymous reviewers and conference organizers for their constructive comments and efforts towards improving this manuscript.

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022)
2. Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Laaksonen, J., Shah, M., Khan, F.S.: Person image synthesis via denoising diffusion model. arXiv preprint arXiv:2211.12500 (2022)
3. Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., Kersting, K.: Sega: Instructing diffusion using semantic dimensions. arXiv preprint arXiv:2301.12247 (2023)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800 (2022)
5. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. arXiv preprint arXiv:2301.13826 (2023)
6. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5343–5353 (2024)
7. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis (2021)
9. 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)
10. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
11. Ge, S., Park, T., Zhu, J.Y., Huang, J.B.: Expressive text-to-image generation with rich text (2023)
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NIPS (2020)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
17. Huang, Z., Wu, T., Jiang, Y., Chan, K.C., Liu, Z.: Reversion: Diffusion-based relation inversion from images. arXiv preprint arXiv:2303.13495 (2023)
18. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7701–7711 (2023)
19. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)

20. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR. pp. 22511–22521 (2023)
21. Liang, Y., Yang, X., Lin, J., Li, H., Xu, X., Chen, Y.: Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching (2023)
22. 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)
23. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII. pp. 423–439. Springer (2022)
24. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2021)
25. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. arXiv preprint arXiv:2211.09794 (2022)
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. 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)
28. Parmar, G., Singh, K.K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. arXiv preprint arXiv:2302.03027 (2023)
29. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
30. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
32. 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)
33. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022)
34. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. PMLR (2015)
35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
36. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems **32** (2019)
37. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)

38. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. arXiv preprint arXiv:2211.12572 (2022)
39. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
40. Wang, L., Yang, S., Liu, S., Chen, Y.c.: Not all steps are created equal: Selective diffusion distillation for image manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7472–7481 (2023)
41. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6232–6242 (2024)
42. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023)
43. Yang, S., Chen, Y., Wang, L., Liu, S., Chen, Y.: Denoising diffusion step-aware models. arXiv preprint arXiv:2310.03337 (2023)
44. Zeng, Y., Lin, Z., Zhang, J., Liu, Q., Collomosse, J., Kuen, J., Patel, V.M.: Scenecomposer: Any-level semantic image synthesis. arXiv preprint arXiv:2211.11742 (2022)
45. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
46. Zhao, M., Bao, F., Li, C., Zhu, J.: Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. arXiv preprint arXiv:2207.06635 (2022)
47. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models (2023)