CoSMOS: Controlling Semantics and Object Counts via Mean-Oriented Steering

Hyoryung Kim, Hyemin Boo, Seunghyeon Lee, Myungjin Lee
Department of Artificial Intelligence,
Ewha Womans University
michellekim0922@ewhain.net, {hyeminb, 2277044, lmjin}@ewha.ac.kr

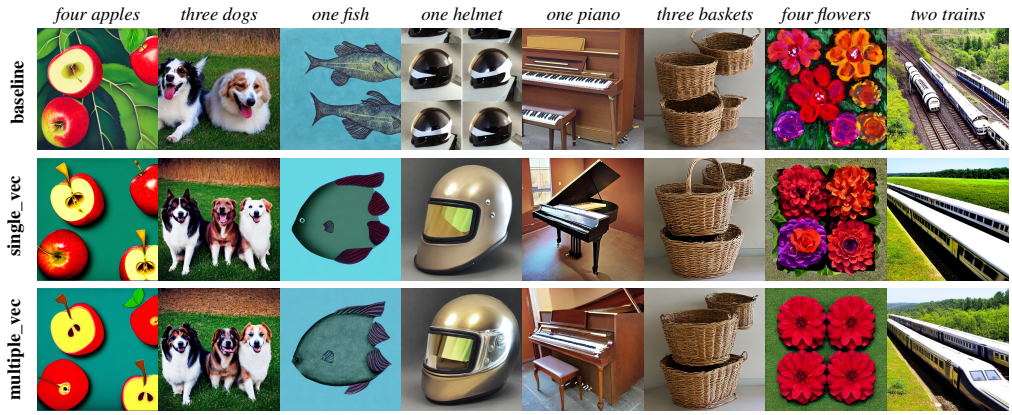


Figure 1: **Comparison of generated images across prompts (columns) and model variants (rows).** Top row shows different prompts; left column shows method types.

Abstract

Text-to-image(T2I) generation models have improved greatly and have been widely used in many applications. However, T2I models still encounter challenges in accurately representing the requested number of objects in the user prompt. In this work, we alleviate this problem by investing an inference-time intervention framework that injects steering vectors into the cross-attention layers of a pre-trained Stable Diffusion model. Our approach modulates the attention query projections using pre-computed direction vectors, without requiring any model fine-tuning. To improve robustness and adaptability, we introduce a dynamic scaling mechanism for the steering strength, computed from the distance and alignment between current latent representations and a reference distribution. We explore both single-vector and step-wise variants of the method and show that our approach enhances control over object count in generated images while preserving image fidelity. This demonstrates that inference-time steering, though simple in implementation, can serve as an effective mechanism for improving numerical fidelity in T2I models. Code is available at: CoSMOS-T2I-Steering.

1 Introduction

Text-to-image generation is the task of generating visual content from natural language prompts. This technology holds significant promise for applications in design, digital art, and visual storytelling. Early approaches, such as Generative Adversarial Networks (GANs) and DALL·E, demonstrated remarkable potential but suffered from critical limitations: GANs often exhibit mode collapse and poor diversity, while DALL·E-based models struggle with fine-grained semantic control, particularly in interpreting complex or quantitative prompts.

With the rise of diffusion models, text-to-image generation has seen notable improvements in both image quality and architectural flexibility. Among them, Stable Diffusion stands out as a widely adopted open-source baseline. However, it still faces a persistent challenge: infidelity with numerical elements in prompts. For example, when asked to generate *"four apples"*, the model frequently produces only *three*, highlighting its inability to accurately control object count. This gap between textual intent and visual realization limits its use in domains where precision matters—such as educational tools, visual instructions, or quantitative illustration.

Interestingly, a similar challenge exists in large language models (LLMs). Li et al. (2023) [8] introduced Inference-Time Intervention (ITI), showing that LLMs may decode correct answers in initial inference steps internally but fail to express them correctly. By steering hidden states using carefully crafted vectors, ITI successfully corrected such failures without retraining the model. Inspired by this, we hypothesize that diffusion models also develop latent awareness of numerical concepts, but this awareness often fails to manifest in the generated output. To address this, we adapt the ITI framework to steer latent representations in Stable Diffusion at inference time, aiming to correct object count mismatches.

Furthermore, prior work Hertz et al. (2022) [3] has shown that object layout—including count—is primarily determined in the early denoising steps of the reverse diffusion process. Guided by this, we limit our intervention to the first 10 steps. Therefore, we apply our steering vectors specifically during these initial steps, targeting cross-attention blocks's query vectors, where queries are related to image latent values and cross-attention blocks are where images and prompts meet.

Our contributions are as follows:

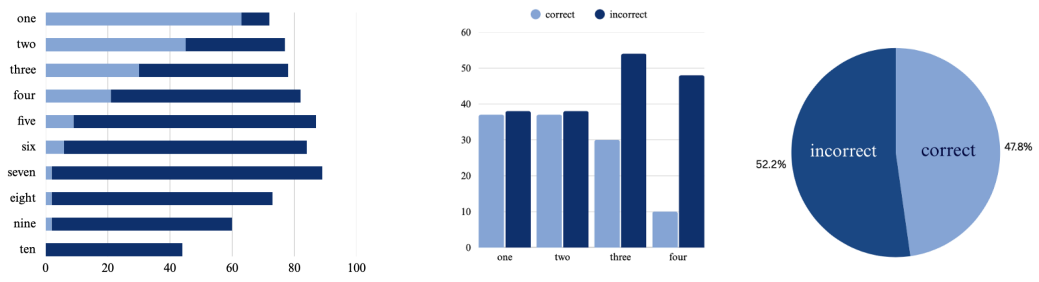
- We propose a training-free, inference-time steering method for aligning object count in generated images with the prompt.
- We introduce a dynamic alpha mechanism that adjusts steering strength based on intermediate activations, enabling adaptive intervention.

2 Related Works

Diffusion Models Diffusion models have emerged as a dominant framework in image synthesis, particularly in text-to-image (T2I) generation, due to their ability to produce images with high fidelity and diversity. These models are based on a two-stage process, a forward diffusion process, which incrementally perturbs data by adding noise, and a reverse denoising process, which reconstructs the original data distribution. The Denoising Diffusion Probabilistic Model (DDPM) [4] first demonstrated the potential of this framework by modeling the forward process as a parameterized Markov chain. However, DDPMs are computationally intensive due to the large number of denoising steps required for sampling. Subsequent methods such as DDIM [13] alleviate this issue by reducing the number of required steps without compromising image quality. Further advancements have been made with Latent Diffusion Models (LDMs) [11], which operate in a compressed latent space rather than pixel space, thereby significantly lowering computational cost and enabling high-resolution image synthesis.

Text-to-Image Models Early approaches to text-to-image (T2I) generation were predominantly based on Generative Adversarial Networks (GANs)[10], which demonstrated initial success in synthesizing images conditioned on textual descriptions. However, GAN-based models suffered from inherent limitations such as training instability and mode collapse, which hindered their scalability and fidelity. To address these limitations, Diffusion-based T2I models have shown remarkable capabilities in translating natural language prompts into high-quality images. Notable examples include stable diffusion[11], Imagen[12], which leverage large-scale datasets, cross-attention mechanisms, and latent space optimization to achieve state-of-the-art performance. These models utilize powerful language encoders and carefully designed denoising architectures to align textual semantics with visual outputs effectively. Despite these successes, T2I diffusion models still struggle with certain aspects of prompt understanding, particularly in maintaining object-level consistency and precise spatial arrangements. Among these challenges, numerical fidelity—generating exactly the number of objects specified in the prompt—remains a significant unsolved problem.

Inference-Time Intervention Li et al.(2025) [8] propose Inference-Time Intervention (ITI), a method to enhance truthfulness in large language models by modifying internal activations during inference. ITI identifies latent directions associated with truthful generations via linear probing and



(a) (b)

Figure 2: **Overall distribution of correct and incorrect counts across different targets.** (a) shows the trend across all targets. (one to ten) As the target object count increased, the model's ability to generate images matching the specified number progressively declined. (b) summarizes the distribution within the internal dataset focused on targets one to four.

shifts model representations toward these directions by adjusting intermediate hidden states. Notably, this approach requires no retraining and only a small number of labeled examples. While ITI was introduced in the context of natural language generation, it demonstrates the broader potential of inference-time latent manipulation for correcting undesirable outputs. Inspired by this perspective, our work extends inference-time steering to vision diffusion models, introducing a dynamic vector-based modulation mechanism that aligns generated images with prompt-specific object semantics.

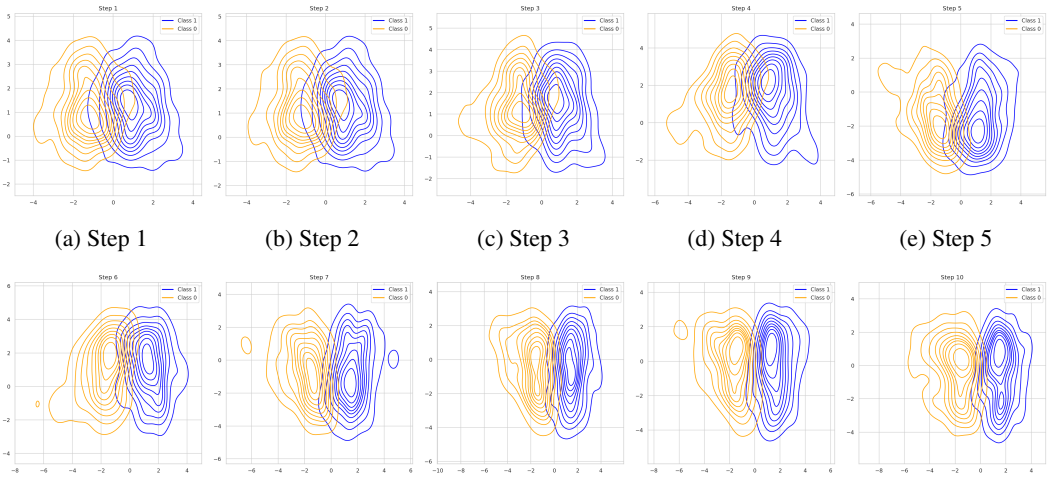
Generating images with accurate object count Recent studies have highlighted the persistent challenge of achieving numerical fidelity in text-to-image (T2I) diffusion models.[2, 1, 6] Empirical evaluations, such as T2I-Compbench[5], consistently show that even state-of-the-art diffusion models struggle to generate the exact number of objects specified in prompts, with performance deteriorating as the requested counts increase. CountGen[1] identifies object-level features during the denoising process to dynamically count object instances and proposes new object locations when under-generation is detected, guiding the diffusion model to achieve accurate object counts without relying on external layouts. CountDiffusion[9] proposes a two-stage framework where an intermediate denoising result is analyzed by a pretrained counting network, and the object count is corrected by modifying attention maps through universal guidance.

In contrast, we propose a lightweight inference-time intervention method based on ITI, which adjusts attention dynamically during generation. This enables effective count control without requiring retraining, external networks, or additional supervision, providing a lightweight practical solution to the numerical reasoning gap in T2I diffusion models.

3 Method

3.1 Setup

Dataset The dataset which is used for steering-vector construction (hereafter referred to as the internal dataset) consists of 292 images, each accompanied by cross-attention hidden states extracted throughout the denoising process. Specifically, the query component of the cross-attention was captured at early denoising steps of the model's attention to the prompt. We hypothesized that although the model successfully encodes prompt information, it often fails to reflect this information accurately in the generated output. To investigate this gap, we focused on the query vectors, which represent the image's hidden state within the UNet architecture. In the cross-attention mechanism, the key and value vectors correspond to the text embeddings derived from the prompt, while the query vectors originate from the latent representation of the evolving image. To provide context for the construction of the internal dataset, we first conducted a preliminary analysis based on 747 generated images, covering prompts requesting object counts ranging from one to ten. As shown in Figure 2a, as the target object count increased, the model's ability to generate images matching the specified number progressively declined. In particular, while lower target counts (*one, two, three, and four*) were often accurately generated, the success rate decreased substantially for target counts more than *five*. This trend highlights the increasing difficulty associated with higher object counts. Based on these observations, we constructed a internal dataset focused on target counts of one to four objects,



(a) Step 1 (b) Step 2 (c) Step 3 (d) Step 4 (e) Step 5

(f) Step 6 (g) Step 7 (h) Step 8 (i) Step 9 (j) Step 10

Figure 3: **KDE visualizations of projected hidden states across denoising steps 1 to 10.** The x and y axes represent projection onto the first and second correct directions, respectively. The orange line represents Class 0 (incorrect images), and the blue line represents Class 1 (correct images). This visualization demonstrates a clear separation between the two classes.

where success rates remain substantial but sufficiently challenging. Figure 2b. The distribution across target counts is as follows: 75 images targeting one object (49.33% success rate), 75 targeting two objects (49.33%), 84 targeting three objects (35.71%), and 58 targeting four objects (17.24%). The overall success rate across the internal dataset is 47.28%, ensuring a reasonably balanced composition despite variations in per-target success rates.

All prompts were generated via GPT-4o prompted with simple *"{count} {object}"* format to focus exclusively on the counting task. Each prompt was paired with fixed random seed to ensure reproducibility. The objects featured in the dataset were drawn from a diverse set of categories, including food, common items, animals, and people. Hidden states of step 1-10 were extracted for all prompts in the internal dataset. For evaluation, an external dataset was constructed using 80 prompts consisting new prompts and random seeds (also via GPT-4o) without hidden-state extraction; these images are not used for steering vector construction.

3.2 Class Seperatibility

To assess the internal validity of our dataset and examine whether the cross-attention query representations contain meaningful class-separating information, we visualized their latent distribution at denoising for step 1 to 10. For each prompt, we extracted the output vectors from the cross-attention modules across all blocks. We then computed mean vectors for each prompt by aggregating block-level vectors.

To quantify class separability, we trained a logistic regression classifier using the extracted vectors and used the learned weight vector as the first projection direction, capturing the most discriminative axis between Class 0 and Class 1. However, as one-dimensional projection was insufficient to visualize all data structure, we introduced a second projection direction. This was computed via principal component analysis (PCA) applied to the residual vectors after removing their component along the first direction, ensuring orthogonality and maximal variance coverage in the remaining subspace.

All hidden vectors were projected onto this 2D space. Figure 3 presents the kernel density estimation (KDE) plots of the projected data, showing a clear separation and mean shift between the two classes. This indicates that class-discriminative features emerge early in the generation process, and this observation served as the empirical foundation for our CoSMOS steering mechanism. This visualization method was inspired by prior work ITI [8], where LLM's hidden states were analyzed via linear probes and 2D projections to assess model internalization of truthfulness. We adapt this analytical technique to the T2I domain to justify our work.

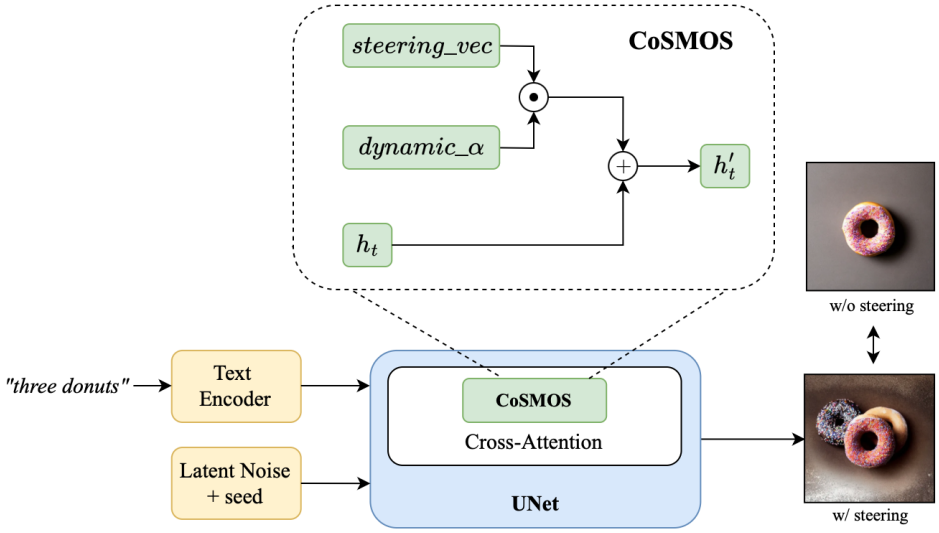


Figure 4: **Overview of our proposed method, CoSMOS.** We dynamically steer the cross-attention during the denoising process based on the discrepancy between the target and current latent mean.

3.3 CoSMOS

Overview This paper proposes an inference-time steering mechanism for diffusion models that enhances numerical fidelity in counting tasks without requiring any fine-tuning. Our approach modifies the cross-attention computation within the denoising U-Net by injecting steering vectors, defined as the difference between latent representations of correct (Class 1) and incorrect (Class 0) images. These vectors are dynamically adapted based on the divergence between the current latent representation and the steering vector during the denoising process, and are injected into the model's attention layers during the early denoising steps, thereby guiding the model to generate images with the correct number of objects. This process consists of two key components: the steering vector $\vec{s}$, which defines the direction of adjustment in the latent space, and dynamic alpha α , which controls the strength of steering at each denoising step based on the model's current state.

Steering Vectors We labeled correct generated images and incorrect ones, and analyze the corresponding attention representations. Upon observing clear separability in their latent distributions, we experimented with various strategies to construct an effective steering vector that captures this distinction. Our final approach defines the steering vector $\vec{s}$ as the difference between the mean hidden states of Class 1 and Class 0. Specifically, each mean, μ_1 and μ_0, is computed by averaging the hidden states over N_1 and N_0 data points from Class 1 and Class 0, respectively. The resulting steering vector is expressed as $\vec{s} = \mu_1 - \mu_0$, and it captures the directional shift required to guide the model's hidden representation from an incorrect generation toward a correct one.

$$\mu_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} h_i^{(1)}, \quad \mu_0 = \frac{1}{N_0} \sum_{j=1}^{N_0} h_j^{(0)} \tag{1}$$

$$\vec{s} = \mu_1 - \mu_0 \tag{2}$$

This paper explores two variants of the steering vector strategy for application during inference.

- **Single Vector**: A single fixed steering vector is computed from the latent activations at denoising step 1, which was empirically selected as it consistently outperformed other early steps in terms of steering effectiveness. This vector is used consistently across 1-10 denoising steps.
- **Multiple Vectors**: A set of steering vectors, each corresponding to a 1-10 denoising step. That is, the vector injected at step t is computed using latent vectors collected specifically at each denoising step.

Dynamic Alpha In our experiments, we observed that applying a static alpha value on the steering vectors often led to unstable and inconsistent results. When the alpha was set too high, the model tended to over-steer, producing unwanted artifacts or generating outputs that deviated from the intended semantics. Conversely, when the alpha was too small, the steering effect was negligible, resulting in minimal or no semantic change in the generated image. These findings highlight a fundamental limitation of static alpha: it does not account for dynamically evolving internal representations during the diffusion process.

To address this issue, we propose a dynamic alpha mechanism that adjusts the steering strength in each denoising step according to the current hidden state. Our method computes alpha by jointly considering the distance between the current hidden representation $\mu_{t,b}$ and the $\mu_{1,b}$, as well as its alignment with the predefined $\vec{s}_{t,b}$ at step t and block b. This design enables sensitive modulation of the steering strength, ensuring stable and effective semantic guidance throughout the generation process. As a result, our approach achieves smoother and more semantically coherent transitions in the output images.

At each denoising step, we compute the error vector $\vec{e}_{t,b}$ between the current hidden mean $\mu_{t,b}$ and the Class 1 mean $\mu_{1,b}$. To quantify directional consistency, we compute the cosine similarity between the steering vector $\vec{s}_{t,b}$ and the error vector $\vec{e}_{t,b}$, referred to as $alignment_{t,b}$. This guides the steering based on directional agreement.

$$\vec{e}_{t,b} = \mu_{1,b} - \mu_{t,b} \tag{3}$$

$$alignment_{t,b} = \frac{\vec{s}_{t,b} \cdot \vec{e}_{t,b}}{\|\vec{s}_{t,b}\| \cdot \|\vec{e}_{t,b}\| + \epsilon} \tag{4}$$

We compute $d_{t,b}$, the L2 distance of the error vector at step t and block b, and normalize it using the magnitude of the steering vector, denoted as $d_{\text{steer},b}$. A scaled exponential function is applied to gradually increase the steering strength based on the semantic gap. The final scaling factor $\alpha_{t,b}$ incorporates both the distance and the alignment. Based on empirical observations, the initial alpha $\alpha_{init}$was set to 100.

$$d_{t,b} = \|\vec{e}_{t,b}\|_2, \qquad d_{\text{steer},b} = \|\mu_{1,b} - \mu_{0,b}\|_2 \tag{5}$$

$$\alpha_{t,b} = \alpha_{init} \cdot \left(1 - e^{-d_{t,b}/d_{\text{steer},b}}\right) \cdot alignment_{t,b} \tag{6}$$

To ensure numerical stability, the dynamic scaling factor $\alpha_{t,b}$ is clipped within a bounded range. The final hidden state h'_t is updated by shifting in the direction of the steering vector $\vec{s}_{t,b}$, scaled by $\alpha_{t,b}$. This allows the model to iteratively move toward the correct distribution based on both semantic deviation and directional consistency.

$$|\alpha_{t,b}| \leq \text{MAX}_{\alpha_{t,b}} \tag{7}$$

$$h'_{t,b} = h_{t,b} + \alpha_{t,b} \cdot \vec{s}_{t,b} \tag{8}$$

At each step t and block b, the dynamic alpha value $\alpha_{t,b}$ is multiplied by the steering vector $\vec{s}_{t,b}$ and added to the original hidden state h_t to produce the updated hidden state h'_t

Intervention to Cross-Attention Previous studies [14, 3] indicate that object layout—particularly object count—is primarily established in the early steps of the reverse diffusion process. Guided by this insight, we restrict our intervention to the first 10 denoising steps. Therefore, we apply the steering vector only during these early iterations to effectively guide semantic formation. To enable intervention during the denoising process, our approach registers hooks on the attention layers, allowing access to and modification of hidden representations. Specifically, the dynamically steered hidden state h'_t is injected into the cross-attention computation only, leaving self-attention and other modules unchanged. This intervention provides semantic guidance while maintaining internal coherence and improving numerical fidelity of the model.

Methods	Internal Dataset			External Dataset		
	ACC ↑	MAE ↓	RMSE ↓	ACC ↑	MAE ↓	RMSE ↓
Baseline	0.4734	1.7381	6.6611	0.5875	2.2375	11.1024
Multiple	0.5102	**1.2959**	**3.5007**	**0.6500**	2.0875	11.1563
Single	**0.5170**	1.5034	6.3749	0.5625	**1.1625**	**3.4695**

Table 1: **Quantitative evaluation on internal and external dataset.**

4 Experiment

This section describes the experimental setup and evaluation framework for analyzing the effectiveness of the proposed steering strategies in T2I object counting fidelity.

Settings All experiments are conducted using Stable Diffusion v1.5 with a fixed number of 50 inference steps and a guidance scale of 7.5, shared across all steering variants including baseline, single-vector, and multiple vector models. Steering interventions are applied only during the first 10 denoising steps, based on the empirical observation that early stages are critical for global object layout and cardinality. The base alpha is set to 100.0. These hyperparameters were chosen based on our experiment to strike a balance between semantic control and image quality.

Evaluation Metrics Object counts are automatically obtained using a LLaVa One-Vision[7] model by prompting each generated image with: *"How many {objects} are in the image? Reply with only a number."* This approach enables evaluation without requiring manual annotation. To assess numerical fidelity, we employ the following metrics: ACC (Accuracy), which indicates whether the predicted count exactly matches the target, and MAE(Mean Absolute Error) and RMSE(Root Mean Squared Error), which measure the magnitude of differences between the predicted and target counts.

4.1 Results

4.1.1 Quantitative Results

Table 1 reports the results on both an internal dataset that was utilized to make a steering vector and an external dataset that contains more diverse and unconstrained prompts. On the internal set, all models perform relatively stably, with the baseline model reaching an accuracy of 0.4734 and the single vector steering method achieving a comparable ACC of 0.5102. Notably, the multiple method significantly reduces the error to 3.5007 and also records the lowest MAE (1.2959), indicating improved alignment even within the seen domain.

In contrast, the external dataset includes a small number of extreme outliers—two or three cases made 100 objects—that substantially inflate the overall MAE and RMSE scores. These outliers typically involve prompts with highly overlapping or densely packed objects, which led to large prediction errors, particularly for the baseline model scoring RMSE (11.1024) and MAE (2.2375). Despite this, the steering-based models remain robust: the multiple vector method lowers the MAE to 2.0875, while the single vector method achieves the lowest RMSE (3.4695) and MAE (1.1625), demonstrating exceptional generalization even under distribution shift. Notably, the multiple method achieves the highest ACC(0.6500). These results suggest that steering—whether applied in a single or multiple method—substantially improves numerical fidelity across both controlled and open-ended prompts without compromising image fidelity.

4.1.2 Qualitative Results

Figure 1 qualitatively compares the generated images across different model variants. Each column corresponds to each prompt and each row corresponds to different steering methods: baseline, single vector, and multiple vector. The baseline failed to capture the correct object count, frequently generating either fewer or more objects than specified. In contrast, the single vector steering model exhibits corrected alignment with the targeted count. The multiple vector steering approach demonstrates better performance than baseline method, successfully generating the desired number

of objects across diverse prompts while preserving any other aspects than numerical fidelity. This indicates that iterative adjustment via dynamic steering enables fine control over numerical constraints in text-to-image synthesis.

5 Conclusion

In this paper, we propose a steering-based approach to improve numerical fidelity in T2I diffusion models. Our method guides the generation process by injecting steering vectors—either the single vector or multiple vector method—into denoising steps. Both steering methods demonstrate consistent improvements over the baseline without additional fine-tuning. Notably, on the external dataset, the multiple vector method increases ACC from 0.5875 to 0.6500. These results suggest that our approach effectively reduces the gap (the model *understands* the number of objects in the prompt, but often fails to *reflect* this numerical accuracy in the generated image) by guiding the model toward better numerical understanding and expression. Qualitative analyzes further support that our method improves alignment with target counts. Overall, our findings highlight the utility of latent-space steering in enabling more accurate and controllable object count in text-to-image synthesis.

6 Future Works

Although our proposed steering vector approach has shown promising capabilities in controlling object count within diffusion-based image generation, several important questions remain open for future exploration. First, extending steering vector construction to a broader and more diverse prompt set is necessary to enhance generalization. Second, a systematic ablation study of blocks, and steps is needed to construct better steering vectors. Third, although steering successfully modulates object count, the underlying factors driving this control remain unclear. In particular, steering sometimes preserves object style while adjusting count, but in other cases causes style changes, or even disrupts originally correct object counts. We hypothesize that steering induces shifts in the hidden state distributions associated with object instances. Further analysis of these distributional shifts could clarify the mechanism behind steering. Finally, evaluating the transferability of steering vectors across different diffusion models would validate generalizability. Through these research directions, we aim to establish a more robust theoretical and empirical framework for controlling object count in image generation, advancing towards generative systems with precise and interpretable quantitative control capabilities.

References

[1] Lital Binyamin et al. Make it count: Text-to-image generation with an accurate number of objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[2] Yuefan Cao, Xuyang Guo, Jiayan Huo, Yingyu Liang, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Zhen Zhuang. Text-to-image diffusion models cannot count, and prompt refinement cannot help. *arXiv preprint arXiv:2503.06884*, 2025.

[3] 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.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[5] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[6] Wonjun Kang et al. Counting guidance for high fidelity text-to-image synthesis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025.

[7] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[8] Kenneth Li et al. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

[9] Yanyu Li, Pencheng Wan, Liang Han, Yaowei Wang, Liqiang Nie, and Min Zhang. Countdiffusion: Text-to-image synthesis with training-free counting-guidance diffusion. *arXiv preprint arXiv:2505.04347*, 2025.

[10] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[11] 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.

[12] 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.

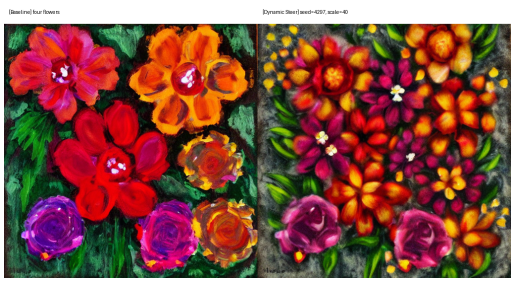
[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[14] Tao Xia, Yudi Zhang, Ting Liu, and Lei Zhang. Consistent image layout editing with diffusion models, 2025. arXiv preprint arXiv:2503.06419.

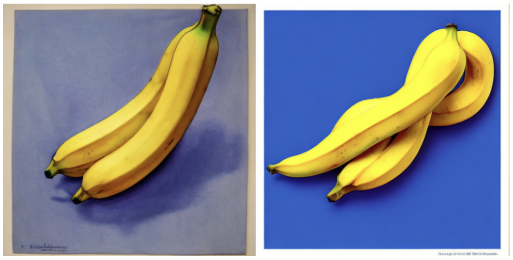
A Appendix

Failure Cases Although the proposed steering mechanism improves numerical fidelity in many cases, several consistent failure patterns have been identified. These are categorized below to clarify current limitations and inform potential enhancements.

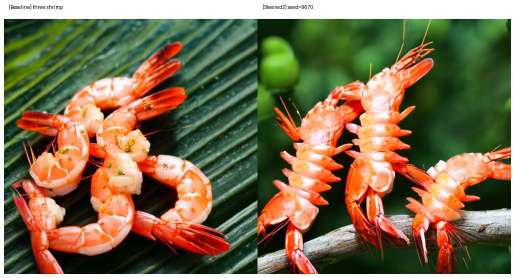
(a) Overgeneration: *one helmet*



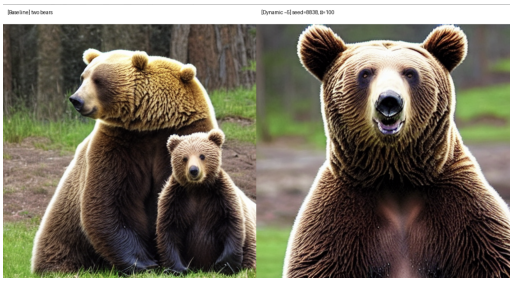
(b) Overgeneration: *four flowers*



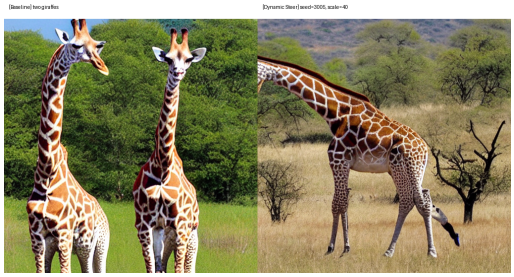
(c) Unreliable Rendering: *three bananas*



(d) Unreliable Rendering: *three shrimps*



(e) Degradation: *two bears*



(f) Degradation: *two giraffes*

Figure 5: **Representative failure cases.** Each case highlights a different type of limitation, including overgeneration, rendering failure, object distortion, and regression from correct to incorrect output.

1. Overgeneration: Unexpected Increase in Object Count

In certain cases, the model generates more objects than intended, despite steering toward a lower count. This behavior may arise when the class-wise latent direction excessively amplifies object-related features, leading to duplication. Such overgeneration is more likely when the object lacks a well-defined or compact representation in the model's latent space.

2. Unreliable Rendering

Steering fails to take effect when the model struggles to generate the object itself. If the object is inconsistently or ambiguously rendered regardless of the prompt, the steering signal derived from latent statistics does not correspond to meaningful semantic directions. This highlights a fundamental dependence on the model's baseline generation capacity.

3. Degradation of Initially Correct Outputs

Instances have been observed where an initially accurate generation becomes incorrect after steering is applied. This regression effect reflects the risk of oversteering, particularly when steering is applied

uniformly without accounting for the confidence or quality of the base output. Such cases emphasize the need for adaptive or selective steering mechanisms that respond to the model's intermediate state.