Progressive Prompt Detailing for Improved Alignment in Text-to-Image Generative Models

Ketan Suhaas Saichandran[1*] Xavier Thomas[1*] Prakhar Kaushik[2] Deepti Ghadiyaram[13†]
[1]Boston University [2]Johns Hopkins University [3]Runway
{ketanss, xthomas, dghadiya}@bu.edu pkaushi1@jhu.edu

Abstract

*Text-to-image generative models often struggle with long prompts detailing complex scenes, diverse objects with distinct visual characteristics and spatial relationships. In this work, we propose SCoPE (**S**cheduled interpolation of **Co**arse-to-fine **P**rompt **E**mbeddings), a training-free method to improve text-to-image alignment by progressively refining the input prompt in a coarse-to-fine-grained manner. Given a detailed input prompt, we first decompose it into multiple sub-prompts which evolve from describing broad scene layout to highly intricate details. During inference, we interpolate between these sub-prompts and thus progressively introduce finer-grained details into the generated image. Our training-free plug-and-play approach significantly enhances prompt alignment, achieves an average improvement of more than **+8** in Visual Question Answering (VQA) scores over the Stable Diffusion baselines on **83%** of the prompts from the GenAI-Bench dataset.*

1. Introduction

Text-to-image diffusion models [18] have made significant strides in generating high-quality generations from textual descriptions. Yet, they struggle to capture intricate details provided in long, detailed prompts describing complex scenes with multiple objects, attributes, and spatial relationships [20, 21]. When processing such prompts, these models often misrepresent spatial relations [3, 22], omit crucial details [9], or entangle distinct concepts [16, 24]. Several reasons contribute to this undesirable behavior. First, the text encoders used to condition the image generation process [14, 15] tend to compress a detailed textual description of varied lengths into a fixed-length representation, potentially leading to concept entanglement or information loss [5]. Second, biases in the pre-training data [19] could be leading to favoring shorter prompts thereby degrading performance on long complex prompts [20].

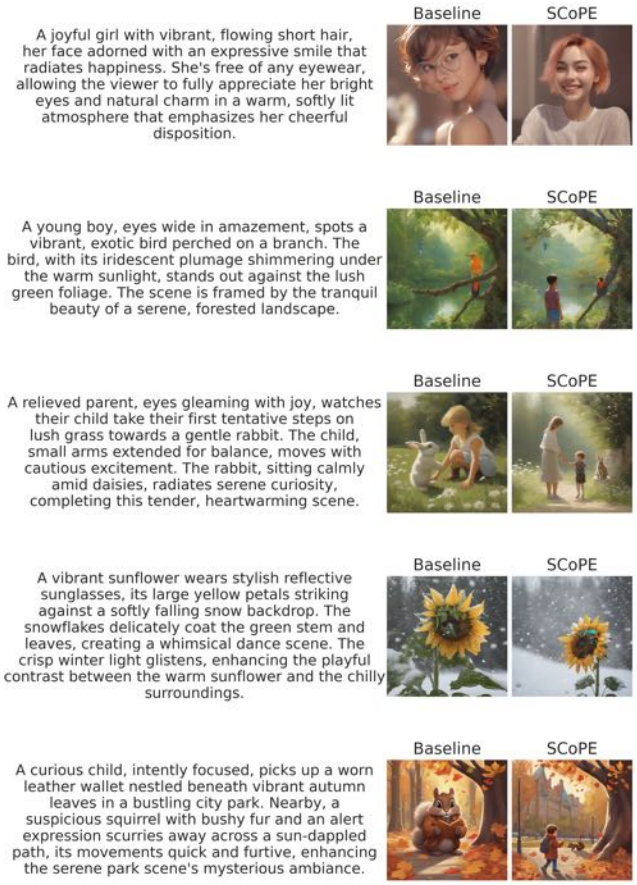


Figure 1. **SCoPE (ours) vs SDXL [12] for long, detailed prompts**. Note how SCoPE (right) captures details mentioned in the prompt better compared to SDXL.

To address this limitation, Zhang et al. [21] extend the context length of the text encoder, allowing for better representation of longer prompts. While effective, this approach requires retraining on large-scale datasets, making it computationally expensive. Another line of work focuses on

*Equal contribution.
†Corresponding author.

mitigating misalignment in the latent space [23] by conditioning on individual concepts sequentially at different stages of the denoising process. However, this method primarily focuses on concept misalignment and entanglement in short prompts and does not focus on addressing the challenges with longer, detailed prompts. In this work, we propose **SCoPE** which stands for **S**cheduled interpolation of **Co**arse-to-fine **P**rompt **E**mbeddings. SCoPE is a training-free approach that improves alignment between the provided (long) prompt and the generated image in diffusion models. Our key idea is to dynamically break down the input prompt into a series of sub-prompts starting from a coarse-grained description that captures the global scene layout to more fine-grained details. We draw inspiration from the findings in Park et al. [11] that diffusion denoising is a progressive coarse-to-fine generation process, where initial timesteps establish low-frequency, global structures, while later steps introduce high-frequency, fine-grained details. Specifically, while all prior methods rely on a single static embedding of the entire input prompt, SCoPE interpolates between progressively detailed prompt embeddings throughout the denoising process, thus generating the global scene layout before gradually introducing finer-grained details. We extensively evaluate SCoPE against several open-sourced models and show that SCoPE improves prompt alignment for long complex prompts (obtained from the GenAI-bench dataset [6]) (see Fig. 1) and achieves a **+8** improvement in VQA-based text-image alignment scores over Stable Diffusion [12, 17] baselines. Notably, SCoPE is both training-free and easily extensible, requires minimal computational overhead of only $+0.7$ seconds per inference on 1 A6000 GPU.

2. Related work

Training on longer texts: Long-CLIP [21] expands the context length of CLIP-based text encoders to handle longer prompts, but requires explicit fine-tuning on long text descriptions. Similarly, Wu et al. [20] improve prompt alignment by fine-tuning both a Large Language Model (LLM) and the diffusion model, leveraging the LLM's semantic comprehension capabilities to better encode the prompt and condition the generation process. However, these approaches rely on high-quality dataset curation and require fine-tuning the diffusion model for prompt alignment. By contrast, SCoPE is training-free, efficient, and greatly improves prompt alignment.

Interpolating text representations: Deckers et al. [2] explore interpolating between two prompt embeddings to control style and content in text-to-image diffusion models. While SCoPE explores a new direction, performing interpolation in a coarse-to-fine-grained manner, progressively refining text guidance throughout the generation process to improve alignment.

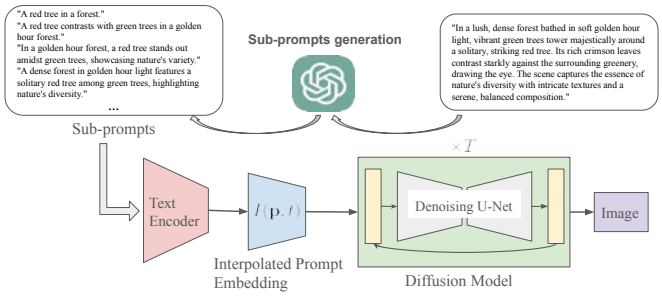


Figure 2. **Training-free approach of SCoPE**, where we first decompose the input prompt into progressively detailed sub-prompts, then interpolates between their embeddings across timesteps, gradually introducing semantic details into the generations.

Idx	Sub-Prompt
1	A cat rides a skateboard on a city street.
2	A cat on a skateboard navigates a city street, balancing amidst people and cars.
3	An earless cat rides a skateboard in a busy street, casting shadows, maintaining balance against a city backdrop.
4	A sleek, earless cat skillfully rides a colorful skateboard on a bustling urban street, highlighted by sunlight shadows, showing agility and balance against a vibrant city backdrop.
5	A sleek, earless cat gracefully rides a colorful skateboard amidst a bustling urban street, effortlessly navigating between pedestrians and vehicles. Sunlight casts dramatic shadows, highlighting the feline's agile, streamlined form as it maintains perfect balance, capturing a sense of motion and boldness against the vibrant city backdrop. †

Table 1. **Progressively detailed sub-prompts derived from the GenAI-Bench prompt *A cat without visible ears is riding.*** † denotes the final prompt used to generate the baseline image. Refer to Fig. 3 for generation results.

Addressing concept misalignment: Zhao et al. [23] highlights how text-to-image diffusion models struggle to accurately compose multiple distinct concepts, and often default to common co-occurring objects from training data. To mitigate this, they introduce concepts sequentially during generation. We build on this intuition and progressively add scene details throughout the denoising process, leading to better alignment with long, complex prompts.

3. Approach

We introduce SCoPE (depicted in Fig. 2), a method for dynamically adjusting text conditioning in diffusion models. First, we describe how sub-prompts are generated from a given input text prompt, each representing a different level of scene granularity (Sec 3.1). Next, we define an interpolation schedule to determine when each sub-prompt has the highest influence during denoising (Sec 3.1). Finally, we describe our interpolation-based text conditioning approach, where the sub-prompts are blended over the denoising steps to guide the image generation process (Sec 3.2).

3.1. Sub-prompt generation and interpolation schedule

Sub-prompt generation. We first use GPT-4o [10] to break down a given prompt into n progressively detailed sub-prompts, each depicting the same scene with increasing

level of detail (an example in Table 1). We then obtain the CLIP embeddings [13] of each sub-prompt such that $\mathbf{p}_1$ corresponds to the embedding of the coarsest prompt and $\mathbf{p}_n$ to the final fine-grained prompt.

Interpolation schedule and interpolation period. During image generation, SCoPE utilizes an interpolated representation (i.e., a weighted sum) of these sub-prompt embeddings for text-conditioning. To determine the timestep where each sub-prompt exerts its maximum influence during denoising, we define an *interpolation schedule* that assigns each sub-prompt $\mathbf{p}_i$ to a specific timestep q_i at which it has the highest influence on image generation. ($i \in \{1, 2, \ldots, n\}$ represents the sub-prompt index). The schedule is initialized at q_1 set to 0, ensuring that the coarsest prompt $\mathbf{p}_1$ guides the early timesteps, where broad scene structures are formed, as noted in [8, 11]. We also define q_n as the *interpolation period* a hyperparameter that determines the timestep up to which interpolation is applied during denoising. We note that interpolation is applied only until timestep q_n, after which $\mathbf{p}_n$ serves as the sole text-conditioning input guiding the diffusion model.

Constructing the interpolation schedule. Instead of uniformly spacing the sub-prompts across the denoising timesteps, we adapt their placement based on the semantic similarity of their embeddings. Specifically, after selecting the hyperparameter q_n, we first set $q_1 = 0$. To determine the remaining timesteps $(q_2, \ldots, q_{n-1})$, we calculate the Euclidean distance between consecutive embeddings, $d_i = \|\mathbf{p}_i - \mathbf{p}_{i-1}\|_2$, and ensure that the ratio $\frac{d_i}{q_i - q_{i-1}}$ remains constant $\forall i \in \{2, 3, \ldots, n\}$. This ensures that semantically similar sub-prompts (i.e. those with smaller Euclidean distances) are assigned timesteps that are closer together, while sub-prompts with greater semantic differences are spaced further apart. We empirically find that this also facilitates a gradual refinement of details throughout the denoising process, which we define next.

3.2. Interpolation-based text conditioning

After defining the interpolation schedule, we use it to apply a Gaussian-based weighting mechanism at each denoising timestep $t \leq q_n$. Specifically, we define a Gaussian of standard deviation σ centered at q_i. This aligns with our motivation where early timesteps benefit from broader, coarse guidance, while later timesteps favor sharper focus on fine-grained details. We define a weight to assign to each prompt embedding $\mathbf{p}_i$ at denoising timestep t as $\alpha_{i,t} = \exp\left(-\frac{(t-q_i)^2}{2\sigma^2}\right)$, where the denominator σ controls the sharpness of the Gaussian function. Following the symmetric decay of the Gaussian, during denoising, at each timestep $t \leq q_n$, weights assigned to earlier sub-prompts gradually decrease, while those for later sub-prompts increase. The weights $\alpha_{i,t}$ are then normalized to obtain $\alpha'_{i,t} = \frac{\alpha_{i,t}}{\sum_{j=1}^{n} \alpha_{j,t}}$. The final text embedding to condition at each timestep is computed as a weighted sum of sub-prompt embeddings, i.e., $I(\mathbf{p}, t) = \|\mathbf{p}_n\| \sum_{i=1}^{n} \alpha'_{i,t} \hat{\mathbf{p}}_i$, where $\hat{\mathbf{p}}_i = \frac{\mathbf{p}_i}{\|\mathbf{p}_i\|}$. The rescaling of magnitude by $\|\mathbf{p}_n\|$ is performed to ensure that the interpolated embedding lies on the hypersphere defined by the CLIP embedding space.

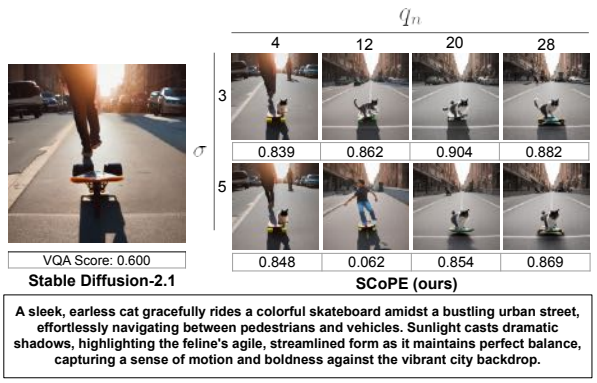


Figure 3. **Effect of σ and q_n on image generation.** The figure illustrates how the variations in standard deviation (σ) and interpolation period (q_n) influence the generated images and their VQAScores. Smaller q_n values (e.g., 4) preserve fine details (e.g., vehicles, pedestrians), while larger q_n values (e.g., 20) emphasize broader structural elements, such as scene composition (e.g., cat, skateboard) and object interactions (e.g., riding). Similarly, smaller σ values lead to images retaining more fine-grained details. See Sec. 4 for more details. The sub-prompts used for this example are provided in Table 1

.

Coarse-to-fine transition in SCoPE. As denoising progresses, influence of coarser sub-prompts at later timesteps decreases. This results in a gradual shift in conditioning from coarse to fine details. The hyperparameter σ further modulates this transition. Higher σ values allows sub-prompts to maintain their influence longer, thereby yielding a more gradual transition. By contrast, lower values of σ makes coarse sub-prompts to lose influence more quickly and resulting in a sharper transition to fine details. We note that for timesteps $t > q_n$, $I(\mathbf{p}, t) = \mathbf{p}_n$, i.e., interpolation is no longer applied, and the model conditions solely on the final fine-grained prompt embedding. This ensures that the image generation process is fully guided by the most detailed prompt, focusing on refining fine-grained details during the later denoising steps, during the fidelity-improvement phase, as discussed in Liu et al. [8]. Thus, the hyperparameter q_n controls when fine-grained details begin to influence the denoising process, determining how early these details start to guide image generation. Fig. 3 shows how σ and q_n impact image generation.

3.3. Evaluation Setup

Dataset: We evaluate our approach using prompts derived from GenAI-Bench [6], which contains 1600 prompts with tags representing spatial relation, counting, negation, etc.

	VQAScore [7]		CLIP-Score [4]	
Model	Mean	Win%	Mean	Win %
SCoPE-v2-1	**87.3**	83.88%	**34.9**	77.56%
SD-v2-1	79.2	-	33.6	-
SCoPE-v1-4	**84.7**	83.44%	**34.5**	74.38%
SD-v1-4	76.6	-	33.3	-
SCoPE-XL	**87.7**	73.00%	**35.3**	65.56%
SDXL	82.9	-	34.6	-

Table 2. **Comparison of mean VQA Scores and CLIP Scores between SCoPE and baseline models.** Win% indicates the percentage of prompts where SCoPE-generated images outperform the baseline. We observe that SCoPE consistently improves over the baselines, regardless of the model.

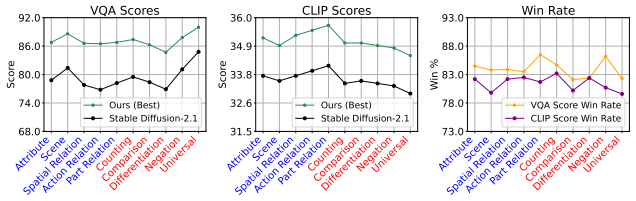


Figure 4. **Comparison between Stable Diffusion-2.1 and SCoPE across different prompt tags in GenAI-Bench [6].** The first five tags (Attribute, Scene, Action Relation, Part Relation, Counting Comparison) are categorized as "Basic," while the remaining tags (Differentiation, Negation, Universal) fall under the "Advanced" category. We observe that SCoPE consistently outperforms the baseline Stable Diffusion-2.1 across both basic and advanced prompt categories.

These prompts have an average token length of 14.8, and often lack fine-grained details. To ensure we test our model on longer, more detailed prompts, we applied the method from Deckers et al. [2] and adopt prompt enhancement, and to target a length of around 50 words, to align within 75 tokens of the CLIP Text encoder. This step results in increasing the average token length to 69.3. Specifically, we use GPT-4o [10] to generate a more detailed version of each prompt. Subsequently, we then use GPT-4o to decompose the enhanced prompt to return four variations, each capturing the same scene at a different level of detail. The prompts used to generate both the enhanced and simplified versions are provided in the appendix. We use the final fine-grained prompt to generate the baseline image generations and to obtain the evaluation scores described next.

Metrics: We evaluate the alignment between generated images and input text prompts (i.e. the fine-grained prompt) using VQAScore [7] and CLIP-Score [4] as our primary evaluation metrics. While CLIP-Score measures the cosine similarity between image and text embeddings, VQAScore [7] uses a Visual Question Answering (VQA) [1] model to produce an alignment score by computing the probability of a "Yes" answer to a simple "Does this figure show {input prompt}?" question. Despite its simplicity, Lin et al. [7] demonstrates that VQAScore outperforms other methods in providing the most reliable text-image alignment scores, particularly for complex prompts. We also report "Win%" as the percentage of prompts where SCoPE-generated images outperform the baseline model.

4. Experiments

We evaluate SCoPE as a plug-and-play approach against Stable Diffusion 1-4 [17], Stable Diffusion-2.1 [17], and SDXL [12]. The total number of inference sampling steps were set to 50 for all models. For each prompt, we generate 8 candidate output images using SCoPE with initial standard deviation $\sigma_0 \in \{3, 5\}$ and interpolation period $q_n \in \{4, 12, 20, 28\}$, as defined in Sec. 3.2. All experiments were carried out on 1 A6000 GPU.

Note on generating candidate outputs. We conduct an empirical study across all 1600 prompts to examine whether VQA Scores [7] correlate with the hyperparameters (σ_0, q_n) and found no clear pattern. In other words, given an input prompt, there is no clear way to predict in advance which setting will yield the most well-aligned generation. To account for this variability, we generate eight candidate outputs per prompt and evaluate them to select the most well-aligned result, described next.

Quantitative results. As shown in Table 2, SCoPE consistently improves text-image alignment, achieving higher VQA Scores and CLIP scores compared to the baselines. For Stable Diffusion-2.1, SCoPE achieves a mean VQAScore of **87.3** across all 1600 prompts derived from GenAI-Bench, outperforming the baseline score of 79.2. We also observe that SCoPE achieves an **83.88**% win rate, indicating that in over 83% of prompts, SCoPE-generated images were more aligned with the input text prompt (as measured by VQAScore). A similar trend is observed for CLIP Score, where SCoPE achieves a mean score of **34.9**, surpassing the baseline score of 33.6, with a 77.56% win rate. For Stable Diffusion 1-4 and SDXL, SCoPE increases VQAScore to **84.7** and **87.7**, with win rates of 83.44% and 73.00%, respectively. In Table 4 we show that SCoPE outperforms the baseline model on varied prompt categories, such as "Spatial Relation," "Counting," "Negation," etc.

5. Conclusion

We propose SCoPE, a simple yet effective, training-free plug-and-play method that improves alignment in text-image generative models, particularly for long and detailed prompts. Our approach offers a lightweight solution that can be seamlessly integrated into existing pipelines. Future work may focus on reducing reliance on candidate outputs and extending applicability to broader generative tasks.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2015. 4

[2] Niklas Deckers, Julia Peters, and Martin Potthast. Manipulating embeddings of stable diffusion prompts. *arXiv preprint arXiv:2308.12059*, 2023. 2, 4

[3] Mohammad Mahdi Derakhshani, Menglin Xia, Harkirat Behl, Cees GM Snoek, and Victor Rühle. Unlocking spatial comprehension in text-to-image diffusion models. *arXiv preprint arXiv:2311.17937*, 2023. 1

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4

[5] Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. *arXiv preprint arXiv:2305.14897*, 2023. 1

[6] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Emily Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Genai-bench: A holistic benchmark for compositional text-to-visual generation. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024. 2, 3, 4

[7] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024. 4

[8] Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Perez-Rua, and Jürgen Schmidhuber. Faster diffusion through temporal attention decomposition. *Transactions on Machine Learning Research*, 2025. 3

[9] Arash Marioriyad, Mohammadali Banayeeanzade, Reza Abbasi, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Attention overlap is responsible for the entity missing problem in text-to-image diffusion models! *arXiv preprint arXiv:2410.20972*, 2024. 1

[10] OpenAI. Gpt-4o, 2024. Version from May 13, 2024. 2, 4

[11] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 2023. 2, 3

[12] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 4

[13] 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*, 2021. 3

[14] 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*, 2021. 1

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 2020. 1

[16] Tanzila Rahman, Shweta Mahajan, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Leonid Sigal. Visual concept-driven image generation with text-to-image diffusion model. *arXiv preprint arXiv:2402.11487*, 2024. 1

[17] 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*, 2022. 2, 4

[18] 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*, 2022. 1

[19] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 2022. 1

[20] Weijia Wu, Zhuang Li, Yefei He, Mike Zheng Shou, Chunhua Shen, Lele Cheng, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Paragraph-to-image generation with information-enriched diffusion model. *arXiv preprint arXiv:2311.14284*, 2023. 1, 2

[21] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, 2024. 1, 2

[22] Gaoyang Zhang, Bingtao Fu, Qingnan Fan, Qi Zhang, Runxing Liu, Hong Gu, Huaqi Zhang, and Xinguo Liu. Compass: Enhancing spatial understanding in text-to-image diffusion models. *arXiv preprint arXiv:2412.13195*, 2024. 1

[23] Juntu Zhao, Junyu Deng, Yixin Ye, Chongxuan Li, Zhijie Deng, and Dequan Wang. Lost in translation: Latent concept misalignment in text-to-image diffusion models. In *European Conference on Computer Vision*, 2024. 2

[24] Chenyi Zhuang, Ying Hu, and Pan Gao. Magnet: We never know how text-to-image diffusion models work, until we learn how vision-language models function. *arXiv preprint arXiv:2409.19967*, 2024. 1

Progressive Prompt Detailing for Improved Alignment in Text-to-Image Generative Models

Supplementary Material

Enhancement Prompt

```
You are an expert at refining prompts
for image generation models.  Your
task is to enhance the given prompt
extensively by adding descriptive
details and quality-improving
elements, while maintaining the
original intent and core concept.
Follow these guidelines:
1.  Preserve the main subject and
action of the original prompt.
2.  Add specific, vivid details to
enhance visual clarity.
3.  Incorporate elements that improve
overall image quality and aesthetics.
4.  Keep the prompt concise and avoid
unnecessary words.
5.  Use modifiers that are appropriate
for the subject matter.
6.  The prompt should contain around
50 words.

Example modifiers (use as reference,
adapt based on some aspect that's
suitable for the original prompt):
- Lighting:  "soft golden hour light",
"dramatic chiaroscuro", "ethereal
glow"
- Composition:  "rule of thirds",
"dynamic perspective", "symmetrical
balance"
- Texture:  "intricate details",
"smooth gradients", "rich textures"
- Color:  "vibrant color palette",
"monochromatic scheme", "complementary
colors"
- Atmosphere:  "misty ambiance",
"serene mood", "energetic atmosphere"
- Technical:  "high resolution",
"photorealistic", "sharp focus"

The enhanced prompt should be short,
concise, direct, avoid unnecessary
words and written as it was a human
expert writing the prompt.
Output only one enhanced prompt
without any additional text or
explanations.
```

Simplification Prompt

```
You are an expert at simplifying
image descriptions.  Your task is to
simplify the description in 4 levels
by removing any unnecessary words
and phrases, while maintaining the
original intent and core concept of
the description.

Follow these guidelines:
1.  Output a python list of 4 such
prompts with increasing levels of
simplification.
2.  Preserve the main subject of the
original description.
3.  Remove all any unnecessary words
and phrases.

Example:

Description:
A pristine sandy beach under a soft
golden hour light, with fine grains of
sand glistening in the warm sunlight.
The shore is framed by gentle waves
lapping at the dunes, creating a
serene mood.  Sparse details of
seashells and driftwood add texture,
inviting tranquility and relaxation.

Output:
['A beach with shiny sand and waves.',
'A beach at sunset with shiny sand.
Waves hit dunes.  Seashells and
driftwood add texture.', 'A beach
at golden hour with glistening sand.
Waves lap at dunes, creating calm.
Seashells and driftwood add texture.',
'A sandy beach at golden hour, with
glistening sand.  Gentle waves lap
at the dunes, creating a serene mood.
Seashells and driftwood add texture.']
```

Tag-wise Scores

Level	Tag	VQA Score		CLIP Score		Win %	
		SCoPE	SD-v2-1	SCoPE	SD-v2-1	VQA	CLIP
Basic	Attribute	0.8683	0.7878	0.3521	0.3367	84.53%	82.22%
	Scene	0.8857	0.8145	0.3490	0.3350	83.78%	79.80%
	Spatial Relation	0.8661	0.7783	0.3535	0.3372	83.87%	82.19%
	Action Relation	0.8649	0.7675	0.3549	0.3387	83.51%	82.46%
	Part Relation	0.8678	0.7820	0.3572	0.3412	86.46%	81.66%
Advanced	Counting	0.8743	0.7949	0.3501	0.3340	84.66%	83.19%
	Comparison	0.8632	0.7845	0.3505	0.3351	82.10%	80.25%
	Differentiation	0.8472	0.7687	0.3489	0.3342	82.43%	82.43%
	Negation	0.8780	0.8105	0.3483	0.3326	86.17%	80.69%
	Universal	0.8997	0.8478	0.3451	0.3302	82.31%	79.59%

Table 3. VQA and CLIP Scores on different tags from the GenAI-Bench [6] dataset for SCoPE and Stable Diffusion-2.1 (SD-v2-1). Absolute improvements are consistent across all categories.

More Qualitative Results

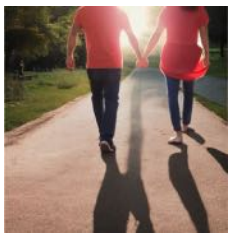
Baseline	SCoPE	
		A determined man dynamically blocks a flying football, standing firmly in front of a small, curious dog. The scene captures the intense action as the football soars through the air, emphasizing the man's protective stance. The vivid setting highlights the contrasting sizes and attributes of the subjects amidst the action
		A single vibrant red rose with delicate petals, elegantly displayed in a crystal-clear vase, positioned gracefully on the right side of a sunlit windowsill. The scene captures soft sunlight filtering through the window, casting gentle shadows and highlighting the rose's vivid hue against the serene backdrop.
		Two vibrant blue and green birds gracefully land side-by-side on an elegant wooden table, surrounded by lush foliage in a serene garden setting, creating a harmonious contrast with their colorful feathers. The scene captures the delicate balance of nature and tranquility, highlighting the beauty of avian life
		A man in a crisp white t-shirt and well-fitted jeans holds hands with a woman in a graceful, flowing red dress, captured in mid-step. Their expressions convey a shared connection as they walk along a sunlit path, surrounded by a soft, warm ambiance.
		In a vibrant, bustling kitchen, chefs of various backgrounds collaborate energetically, each meticulously crafting their unique dishes. Stainless steel surfaces gleam under warm ambient lighting, as fragrant ingredients and colorful spices are artfully arranged. Pans sizzle, knives rhythmically chop, and teamwork creates a harmonious culinary symphony.
		In a serene pond, two ducks gracefully glide over the water. The larger duck boasts a vivid green head contrasted by a smaller, fully brown companion. Rippling reflections dance on the surface, highlighting their distinct features and size difference. Sunlight filters through lush surroundings, casting a warm, inviting glow on the peaceful scene.

Figure 5. More examples to compare SCoPE generated images and the images generated from the baseline Stable Diffusion-2.1.