The Supplementary Materials for
Dynamic Prompt Optimizing for Text-to-Image Generation

Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen and Qing Yang

In this appendix, we provide additional information and materials to complement our research, including training samples, more qualitative results, additional experimental details, and discussion.

A. Examples of training data

We utilize a diverse range of text-image pairs sourced from public datasets and online communities. As shown in Fig. A4, we present some prompts that are included in our training data. These prompts have undergone filtration and construction following the automated process described in Sec. 3.3 of the main manuscript. The short prompts $\mathbf{s}$ primarily describe the subject matter of the images, while the modifiers (highlighted in gray) provide additional details and enhance the aesthetic appeal of the images. In the figure, the term "Aes" denotes the aesthetic score, and "CLIP" quantifies the semantic relevance of the generated image to the short prompt. We can see that the generated images $\mathbf{I}'$ corresponding to the original prompt $\mathbf{s}'$ are more visually effective than the generated images $\mathbf{I}$ corresponding to the short prompt $\mathbf{s}$.

Lexica.art	Aes	CLIP
Short Prompt	5.58	0.28
+ "artstation"	5.83	0.26
+ "concept art"	5.68	0.30
+ "digital painting"	5.79	0.30
+ "sharp focus"	5.60	0.28
+ "highly detailed"	5.64	0.29

Table A1. The effect of different words on generating images.

B. More detailed statistical analysis

Fig. A1 indicates a predominance of shorter token sequences in model predictions, implying that adding a few modifiers can significantly enhance an image's visual appeal without altering the original prompt's meaning. Fig. 5 (b-d) of the main manuscript show frequently generated modifiers, most of which are trends, styles, and texture terms. We also conduct experiments to analyze word impact. As shown in Tab. A1, "artstation" boosts Aesthetic scores at the cost of text-image similarity, whereas styles and texture modifiers slightly increase Aesthetic scores while preserving alignment.

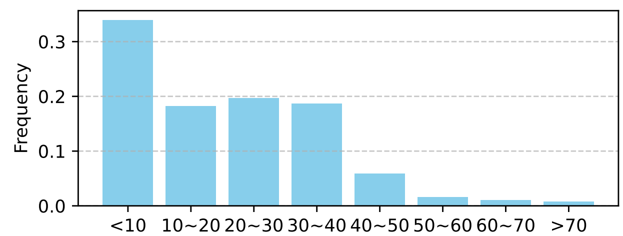


Figure A1. Frequency of the number of predicted word tokens.

C. Enhanced text encoder for DF-prompt

In Stable Diffusion, the text encoder is modified to achieve fine-grained control over the generated effects. These modifications involve two key aspects:

- We introduce weights for each word embedding, representing the impact of a word or phrase on the resulting image. To accomplish this, we apply a weighting operation to each word's embedding by multiplying it with a specific weight. Subsequently, we normalize the entire set of text embeddings, ensuring that the overall mean value remains consistent with the original text embeddings. This normalization step is crucial for maintaining numerical stability. Our technique yields results similar to the existing prompt weighting method (Fig. A2 (b)) but having dynamic time-range control. The pseudo-code for weighting tokens is below.

```
# Given text_embs:[77x768], weights:[77,]
previous_mean = text_embs.mean() # float
text_embs *= weights
current_mean = text_embs.mean()
text_embs *= previous_mean / current_mean
```

Listing 1. Python pseudo code for weighting tokens.

- The injection time steps are regulated using a dictionary. This dictionary maps each word or phrase to a designated time step, which determines when to initiate and conclude the injection of that specific word or phrase during the image generation process. By manipulating the time steps in

Method	Training		Inference (per prompt)		T2I Pipeline (per image)	
	Stage 1	Stage 2	Ours	Promptist	Vanilla SD	Dynamic SD
GPU Times	18 hours	3 days	0.73s	0.69s	5.64s	5.71s

Table A2. Experiment on an A800 (80GB) GPU.

the dictionary, precise control over the duration of different concepts within the generated image can be achieved. These modifications empower the text encoder to exert more precise control over the effects within the Stable Diffusion framework. As a result, more personalized and user-specific image-generation outcomes can be attained.

Method (+"DSLR")	FID (↓)
Promptist	70.80
PAE (Ours)	**69.84**

Table A3. Quantitative comparison of image quality between our method and Promptist, measured using the FID metric.

D. More experimental details

For the evaluation process, we use a maximum new token length of 75 for all evaluated models. We use a temperature of 0.9 during the evaluation and apply a top-k sampling strategy with a k-value of 200. To ensure consistency, we use the same seed in all quantitative evaluation experiments.

E. More qualitative results

In this section, we present more qualitative results, as depicted in Figs. A5 and A6. We compare the images $\mathbf{I}^{\mathrm{DFP}}$ generated using DF-Prompts with the images $\mathbf{I}$ generated using the short prompts. For example, in Fig. A5, we observe that the images corresponding to DF-Prompts, $\mathbf{I}^{\mathrm{DFP}}$, exhibit more vibrant details and aesthetically pleasing color combinations compared to the images $\mathbf{I}$ generated from the short prompts. Some specific examples include "symmetry!! portrait of a warrior transformers robot", "a symmetrical portrait of a beautiful menacing lilith", "commission of a fit male anthro albino lion holding a sword" and "glowwave portrait of dark batman from overwatch". We ensure fairness and consistency by generating the columns corresponding to $\mathbf{I}$ and $\mathbf{I}^{\mathrm{DFP}}$ using the same seed.

Empirical evidence shows that our method not only creates aesthetically pleasing images but also caters precisely to user queries, such as achieving photorealism with "DSLR" or creating 3D-rendered effects with "3D blender" in user prompts (Fig. A2 (a)). Our method shows adaptability when integrating detailed modifiers like "DSLR" and achieves competitive Frechet Inception Distance (FID) [1] (Tab. A3). This adaptability is critical in practical applications.

The time cost of each stage is shown in Tab. A2. As for inference, the average time is marginally higher than that of Promptist (+0.04 s). Moreover, our Dynamic Stable Diffusion (Dynamic SD) method is slightly slower than the Vanilla SD method, but the difference is minimal.

F. Discussion

The significant enhancement in image quality and text alignment observed in Fig. 3 of the main manuscript for the case "cats in suits smoking cigars together" can be attributed to our model's reward mechanism. Specifically, we incorporate the Aes score to encourage actions that improve aesthetic features and the CLIP score to ensure semantic coherence. Additionally, our reward function introduces the PickScore, which allows for more diverse prompt modifiers and ultimately leads to improved image quality. In Fig. 3 of the main manuscript, the inclusion of new semantic elements like "on a ship deck" alongside other modifiers contributes significantly to the visual appeal of the generated output.

Differences among PAE, Promptist, *hugging face weighting prompt method* (WP)[1] lie in that Promptist focuses solely on prompt expansion, while WP manipulates the likelihood of certain phrases appearing in images by artificially setting their weights. PAE, on the other hand, innovatively introduces dynamic prompts, and dynamically adjusts the weights of different phrases during various stages of image denoising, thus achieving more granular control over the image generation process. Additionally, PAE introduces a richer set of reward metrics (aligning closely with user preferences), without the need for manual intervention, resulting in visually striking and semantically consistent images.

As shown in Figs. A5 and A6, the generated image $\mathbf{I}^{\mathrm{DFP}}$ maintains the identity consistency of the image $\mathbf{I}$ produced by the short prompt $\mathbf{s}$ when using the same seed. Meanwhile, it incorporates additional image details that enhance visual appeal. This is evident in Fig. A5 with the example of a "commission of a fit male anthro albino lion holding a sword," and in Fig. A6 with "Grimes with elf ears." This capability can be further developed to ensure consistent role generation. To further enhance our model, it is advantageous to incorporate more comprehensive reward considerations. For instance, evaluating generated images based

[1] https://huggingface.co/docs/diffusers/using-diffusers/weighted_prompts

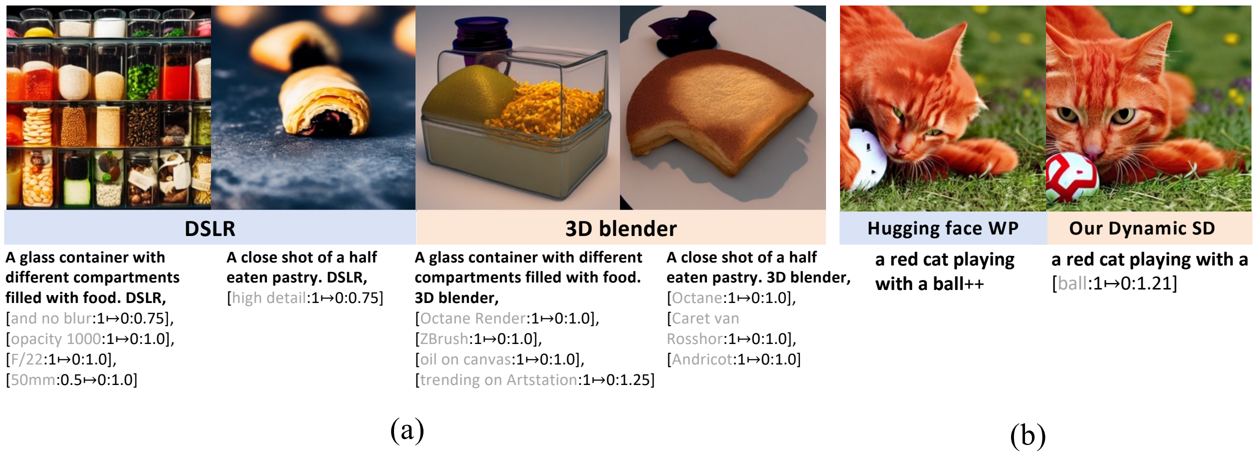


Figure A2. (a) Examples of practicability. (b) Weight methods.

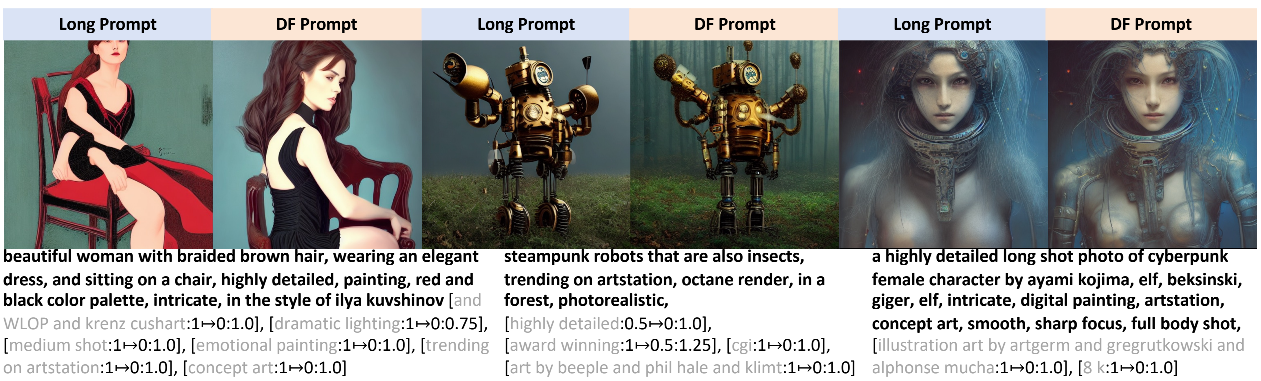


Figure A3. The input long prompts are in bold.

on factors such as high resolution and proportional composition can contribute to their overall quality and realism. Furthermore, to address issues such as attribute leakage and missing objects observed in the original Stable Diffusion method, advanced control techniques can be explored. One potential approach involves incorporating control attention maps into the action space. By selectively directing attention to specific regions in the input image, the model gains finer control over the generation process. Consequently, issues related to attribute leakage can be mitigated, and the preservation of important elements can be ensured. By exploring these possibilities and developing more sophisticated control mechanisms, we can enhance the capabilities of our model and overcome the limitations observed in its current implementation.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 2

Figure A4. Some examples of the training data.

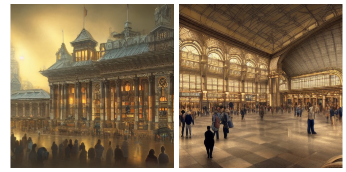
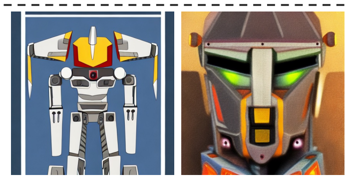
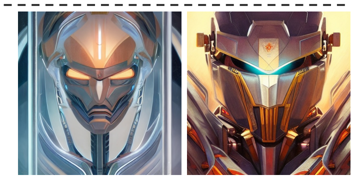
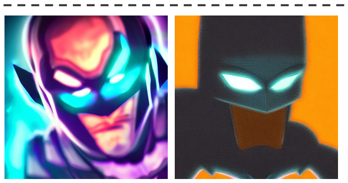
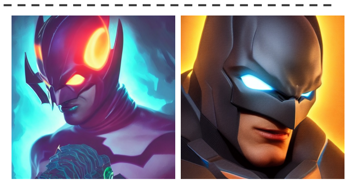
Short Prompt $\mathbf{s}$	Generated Image $\mathbf{I}$	DF Prompt $\mathbf{s_{DFP}}$	Generated Image $\mathbf{I_{DFP}}$
an ultradetailed render of a grand train station		**an ultradetailed render of a grand train station**, [*a large city*:1↦0:1.25], [*many bridges*:1↦0:1.0], [*airbrushed*:1↦0:1.0], [*digital painting*:1↦0:1.0], [*digital painting*:1↦0:0.75], [*trending on artstation*:1↦0:1.0]	
toronto viewed from a distance in the atacama desert		**toronto viewed from a distance in the atacama desert**, [*intricate*:1↦0:1.0], [*elegant*:1↦0:1.0], [*highly detailed*:1↦0:1.0], [*digital painting*:1↦0:1.0], [*artstation*:1↦0:1.0], [*concept art*:1↦0:1.0], [*sharp focus*:1↦0:1.0],[*illustration*:1↦0:0.75], [*by justin gerard and artgerm*:1↦0:1.0], [*8 k*:1↦0:1.0]	
symmetry!! portrait of a warrior transformers robot		**symmetry!! portrait of a warrior transformers robot**, [*intricate*:1↦0:1.0], [*elegant*:1↦0:1.0], [*highly detailed*:0.5↦0:0.75], [*digital painting*:1↦0:1.0],[*artstation*:1↦0:1.0], [*concept art*:0.5↦0:1.0],[*smooth*:1↦0:0.75], [*sharp focus*:0.5↦0:1.0],[*illustration*:1↦0.5:1.25], [*art by artgerm and greg rutkowski and alphonse mucha*:1↦0:1.0], [*8 k*:1↦0:1.0]	
a symmetrical portrait of a beautiful menacing lilith		**a symmetrical portrait of a beautiful menacing lilith**, [*art by artgerm and greg rutkowski and alphonse mucha*:1↦0:1.0], [*volumetric lighting*:0.5↦0:1.0], [*octane*:1↦0:1.0], [*4 k resolution*:1↦0:1.0], [*trending on artstation*:1↦0:1.0], [*masterpiece*:1↦0:1.25]	
commission of a fit male anthro albino lion holding a sword		**commission of a fit male anthro albino lion holding a sword**, [*dnd*:1↦0:1.0], [*face*:0.5↦0:1.25], [*fantasy*:1↦0:1.0], [*intricate*:1↦0:1.0], [*elegant*:1↦0:1.0], [*highly detailed*:1↦0:1.0], [*digital painting*:1↦0:0.75], [*artstation*:1↦0:1.0], [*concept art*:1↦0:0.75], [*smooth*:1↦0:1.0], [*sharp focus*:1↦0:1.0], [*illustration*:1↦0:1.0], [*art by artgerm and greg rutkowski and alphonse mucha*:1↦0:1.0]	
glowwave portrait of dark batman from overwatch		**glowwave portrait of dark batman from overwatch**, [*!! and cthulhu*:1↦0:1.0], [*intricate*:0.5↦0:1.25], [*elegant*:1↦0.5:1.25], [*highly detailed*:0.5↦0:1.0], [*digital painting*:1↦0:1.0], [*artstation*:1↦0:1.0], [*concept art*:1↦0:1.0], [*smooth*:0.5↦0:1.0], [*sharp focus*:0.5↦0:1.0], [*illustration*:1↦0:1.0], [*art by artgerm and greg rutkowski and alphonse mucha*:1↦0:1.0]	
charming muscular gnome engineer		**charming muscular gnome engineer**, [*art by lois van baarle and loish and ross tran and rossdraws and sam yang and samdoesarts and artgerm and saruei and disney*:1↦0:1.0], [*digital art*:1↦0:1.0], [*highly detailed*:1↦0:1.0], [*intricate*:1↦0:1.0], [*sharp focus*:1↦0:0.75], [*trending on artstation hq*:1↦0:1.0], [*deviantart*:1↦0:1.0], [*unreal engine 5*:1↦0:1.0], [*4 k uhd image*:1↦0:1.0]	
floating island with new york city in the sky		**floating island with new york city in the sky**, [*by greg rutkowski*:1↦0:1.0], [*digital art*:0.5↦0:1.0], [*realistic painting*:1↦0:1.0], [*fantasy*:1↦0:1.0],[*very detailed*:1↦0:0.75], [*trending on artstation*:1↦0:1.0]	

Figure A5. More examples of the generated images.

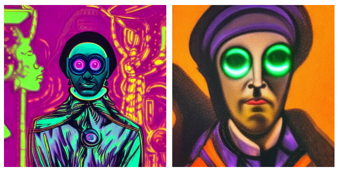
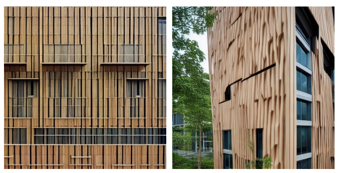
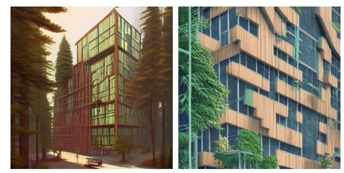
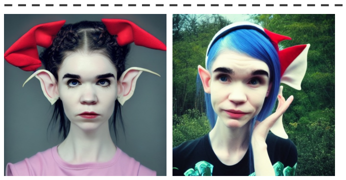
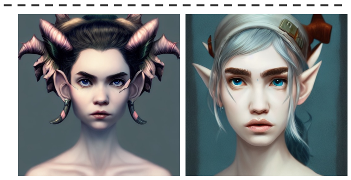
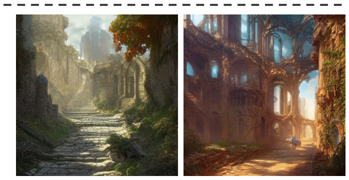
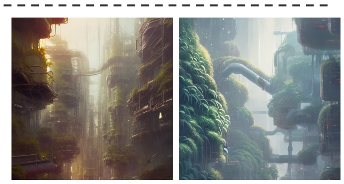
Short Prompt s	Generated Image I	DF Prompt s_{DFP}	Generated Image I_{DFP}
beautiful cottagecore Black Cat chasing butterflies in a dense flower garden		**beautiful cottagecore Black Cat chasing butterflies in a dense flower garden**, [*By Makoto Shinkai*:0.5↦0:1.0], [*Stanley Artgerm Lau*:1↦0:1.0], [*WLOP*:1↦0:1.0], [*Rossdraws*:1↦0:1.0], [*James Jean*:1↦0:1.0]	
a baroque neoclassicist close - up portrait of a colorful retrofuturistic blacklight uv cyborg scientist wizard with glowing eyes		**a baroque neoclassicist close - up portrait of a colorful retrofuturistic blacklight uv cyborg scientist wizard with glowing eyes**, [*glowing fog in the background*:1↦0:1.0], [*renaissance portrait painting*:1↦0:0.75], [*highly detailed science fiction painting by norman rockwell*:1↦0:1.0], [*gustave*:0.5↦0:1.0]	
futuristic timber building facade with wi dows and vegetation		**futuristic timber building facade with wi dows and vegetation** [*by Michael Whelan and Tomer Hanuka*:1↦0:1.0], [*hyperdetailed*:1↦0:1.0], [*artstation*:1↦0:0.75], [*cgsociety*:1↦0:1.0], [*8 k*:1↦0:0.75]	
grimes with elf ears		**grimes with elf ears**, [*intricate*:1↦0:1.0], [*elegant*:1↦0:1.0], [*highly detailed*:1↦0:0.75], [*digital painting*:1↦0:1.0], [*artstation*:1↦0:1.0], [*concept art*:1↦0:1.0], [*smooth*:0.5↦0:1.25],[*sharp focus*:0.5↦0:1.0], [*illustration*:1↦0:1.0]	
a vibrant emotional digital 3 d cg of stone pathway to dystopian post - apocalyptic abandoned castle		**a vibrant emotional digital 3 d cg of stone pathway to dystopian post - apocalyptic abandoned castle**, [*intricate*:1↦0:1.25], [*elegant*:1↦0:1.0], [*highly detailed*:1↦0:1.0], [*digital painting*:1↦0:0.75], [*artstation*:1↦0:1.0], [*concept art*:1↦0:1.0], [*smooth*:1↦0:1.0], [*sharp focus*:1↦0:1.0], [*illustration*:1↦0:1.0], [*art by artgerm and greg rutkowski and alphonse mucha*:1↦0:1.0], [*8 k:*1↦0:1.0]	
wonderdream faeries lady feather wing digital art painting fantasy bloom vibrant		**wonderdream faeries lady feather wing digital art painting fantasy bloom vibrant**, [*wlop*:0.5↦0:1.0], [*greg rutkowski*:1↦0:1.0], [*artgerm*:1↦0:1.25], [*alphonse mucha*:1↦0:1.0], [*beautiful dynamic dramatic dark moody lighting*:1↦0:1.0], [*shadows*:1↦0:1.0], [*cinematic atmosphere*:1↦0:1.0], [*artstation*:1↦0:0.75], [*octane render*:1↦0:1.0], [*8 k*:0.5↦0:1.0], [*masterpiece*:0.5↦0:0.75], [*concept art*:0.5↦0:1.0]	
plants growing out of old rusty pipes in a futuristic city		**plants growing out of old rusty pipes in a futuristic city**, [*By Makoto Shinkai*:1↦0:0.75], [*Stanley Artgerm Lau*:1↦0:1.0], [*WLOP*:1↦0:1.0], [*Rossdraws*:1↦0:1.0], [*James Jean*:1↦0:0.75]	
a painting of a person riding a bike down a dirt road surrounded by vibrant colorful trees		**a painting of a person riding a bike down a dirt road surrounded by vibrant colorful trees**, [*hyperdetailed*:1↦0:1.0], [*artstation*:0.5↦0:1.0], [*cgsociety*:1↦0:1.0], [*8 k*:0.5↦0:1.0]	

Figure A6. More examples of the generated images.