Draw Your Mind: Personalized Generation via Condition-Level Modeling in Text-to-Image Diffusion Models

Supplementary material

7. Detailed effectiveness comparison

We conduct a comparative analysis of performance with the baseline, which was not covered in Section 5.4. As shown in Table 2, FABRIC achieves satisfactory CLIP scores through direct image guidance, but operates only with Stable Diffusion V1 and V2. TV, even with an upgraded GPT, shows the lowest performance on PIP among baselines, and does not exceed the target and history performance of the original T2I models on ML. This limitation is likely due to its reliance on prompt-level modeling, which relies on a limited number of available references. In contrast, PMG employs keyword-centered modeling, which results in some improvement in the historical CLIP score on ML, but it operates only in specific domains such as movies, with suboptimal performance.

8. Impact of sampling method on CLIP score

Table 4 represents the variation of CLIP score as the sampling ratio changes. Coreset sampling achieves the best performance on PIP, while all methods show similar performance on ML. However, since coreset sampling achieves the highest Text align in Figure 3 of the main paper, it indicates that this method provides better personalization on ML.

CLIP score ↑	PIP				ML			
	1%	10%	25%	100%	1%	10%	25%	100%
Random	15.79	15.86	15.83	**15.85**	**13.06**	**13.08**	**13.05**	**13.06**
Uniform	15.84	15.84	15.87		13.05	**13.08**	**13.05**	
Coreset	**16.04**	**16.03**	**16.03**		13.04	13.07	**13.05**	

Table 4. Sampling performance of personalization on CLIP score for different methods using Stable Diffusion V1 and OpenCLIP ViT-L. Random selects randomly and uniform choices evenly.

9. Impact of personalization degree

Figure 8 shows the average performance on both target and history prompts under varying personalization degrees α. To ensure the target's originality, we considered a scale range of 0.1-0.3 for α. In the PIP dataset, performance increases as α grows. In contrast on ML, the CLIP score and Text align intersect at different α, as the accumulation of product details to the image as Text align increases. Based on these experiments, we determined the optimal value of α for DrUM.

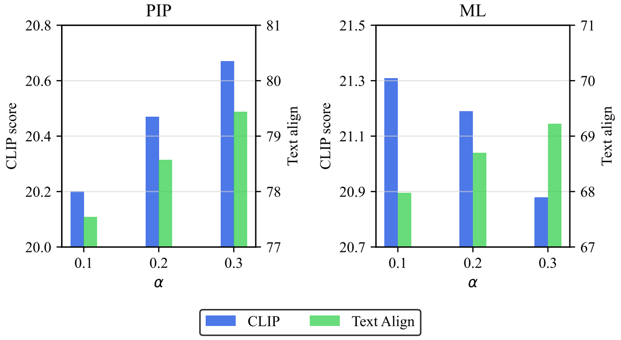


Figure 8. Average performance between target and history for different personalization degree α using Stable Diffusion V1 and OpenCLIP ViT-L.

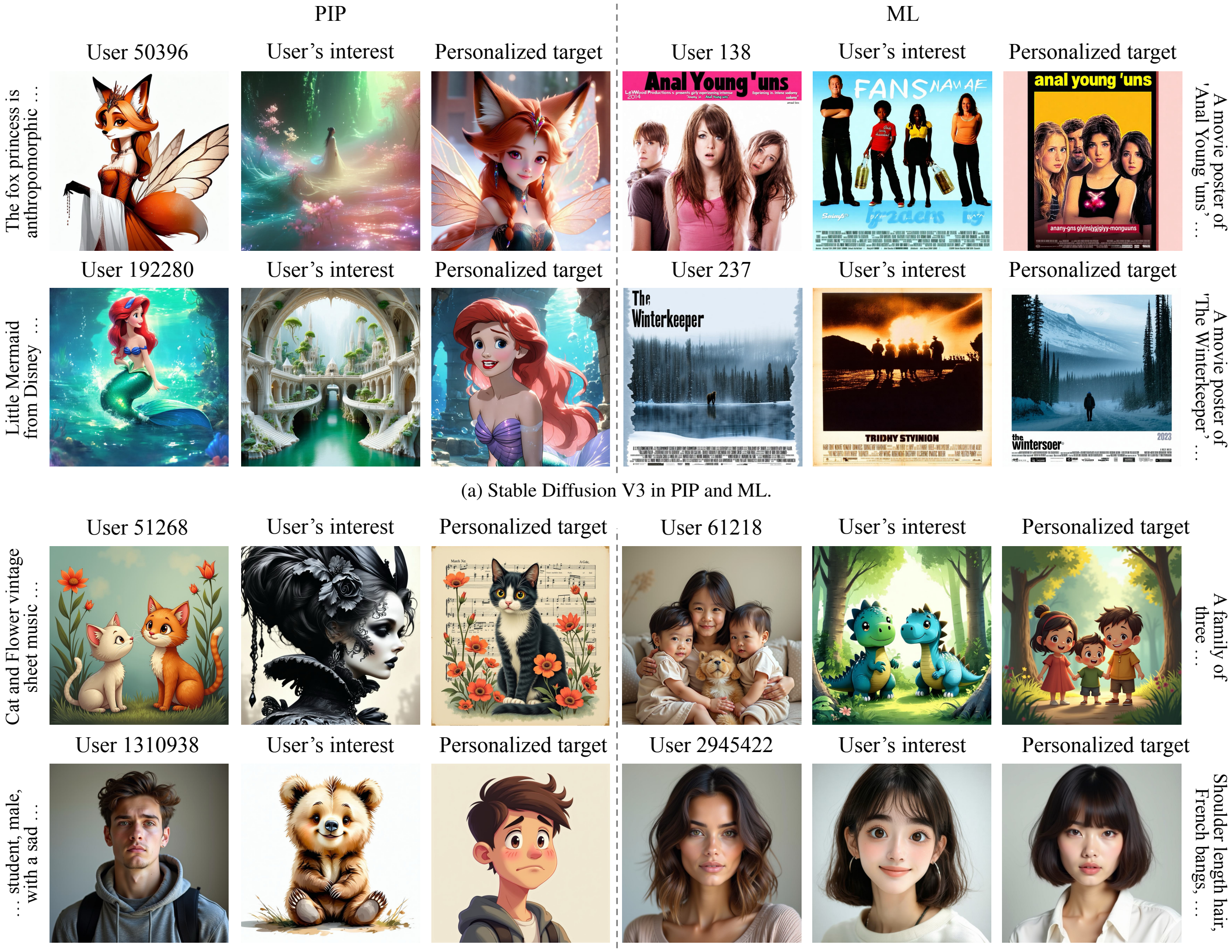
CLIP score ↑	PIP			ML		
	Target	History	Imp	Target	History	Imp
-	20.69	10.99	-	32.41	11.78	-
U	25.49	**15.85**	+34.88%	**29.56**	13.06	+0.46%
L	**25.57**	15.78	+34.73%	28.39	12.48	-3.78%
$U + L$	25.52	15.79	+34.66%	28.41	**13.14**	-1.05%

(a) CLIP score

Text align ↑	PIP			ML		
	Target	History	Imp	Target	History	Imp
-	100.00	52.60	-	100.00	34.16	-
U	98.12	**60.75**	+6.80%	99.59	**36.36**	+3.02%
L	**98.14**	60.66	+6.72%	**99.62**	36.28	+2.91%
$U + L$	98.14	60.73	+6.79%	99.58	36.33	+2.97%

(b) Text align

Table 5. Impact of input queries without coreset sampling on CLIP score and Text align using Stable Diffusion V1 and OpenCLIP ViT-L. Unconditional text embeddings U and learnable embeddings L are considered.



(a) Stable Diffusion V3 in PIP and ML.

(b) Flux in PIP.

Figure 9. Visualization of DrUM using *multiple prompts*. User's interest illustrates the results for $\alpha = 1$ using only reference prompts.

10. Impact of input queries

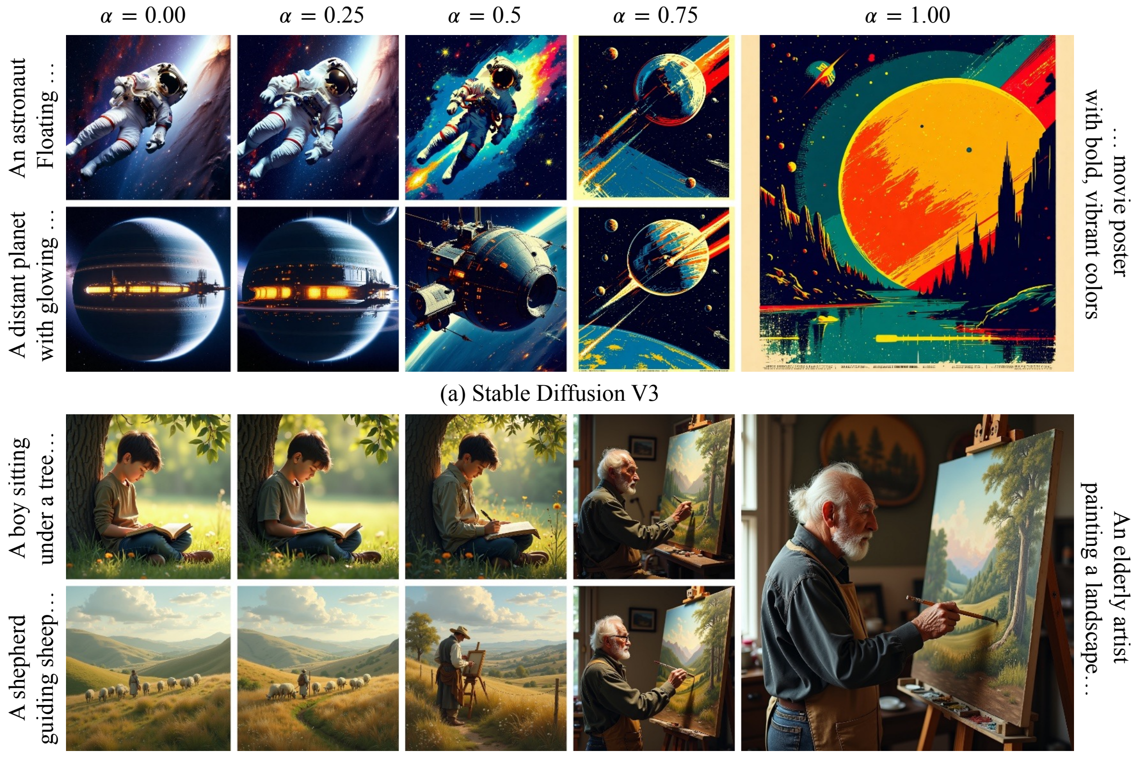
Table 5 presents the performance variations of input queries. Using only unconditional text embeddings achieves the highest performance, while the two cases of learnable embeddings result in decrease CLIP score. This indicates that unconditional text embeddings effectively represents the central distribution of the text encoder for guidance mechanism on DrUM.

11. Additional visualizations of DrUM

Figure 9 presents additional qualitative results highlighting DrUM's content generation capabilities. The figure consists of three main components: Target, user's interest, and personalized results. Here, the user's interest is generated by setting $\alpha = 1$ and using only the reference prompts, enabling an intuitive understanding of how the target is combined with user preferences. In Figure 9a, we utilize Stable Diffusion V3 to observe how the interest's details are reflected in both PIP and ML. Furthermore, Figure 9b shows that the user's interests can cause significant changes in style and texture while maintaining originality. In summary, DrUM estimates the user's interests at the condition-level and effectively integrates them into the synthesized results.

12. Personalized degree in additional scenario

In Section 5.5, the effectiveness of visual changes based on DrUM's personalized degree was demonstrated. Figure 10 further verifies its robustness between dual prompts and a single prompt. In (a), the astronaut and planet effectively reflect the color scheme of the movie poster, while in (b), individuals precisely blend the view and occupation.


α = 0.00
α = 0.25
α = 0.5
α = 0.75
α = 1.00
An astronaut Floating ...
A distant planet with glowing ...
... movie poster with bold, vibrant colors
A boy sitting under a tree...
A shepherd guiding sheep...
An elderly artist painting a landscape...

(a) Stable Diffusion V3

(b) Flux

Figure 10. Impact of personalization degree α between dual prompts and single prompt.

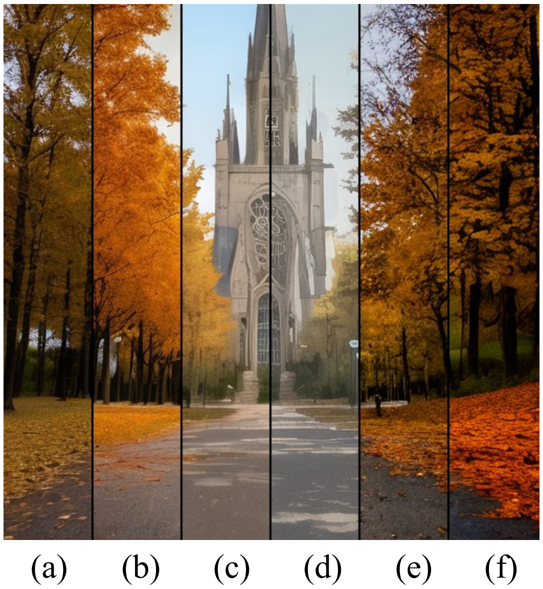


Figure 11. DrUM's adaptability across six open-source T2I models. The target prompt is ”A photo in front of the cathedral on the forest” and the reference prompt is ”A picture of fall”. (a) Realistic Vision V5.1. (b) Deliberate V2. (c) Anything V5. (d) Abyss Orange Mix2. (e) Stable Diffusion V1.5. (f) Open Journey V4.

13. Adaptability

As introduced in Section 5.1, DrUM can be applied without additional fine-tuning when leveraging the same text-encoder. This has been demonstrated through applications to various foundation T2I models, and we further visualize its ability to achieve the same level of personalization performance. Figure 11 presents the personalization results across different models using OpenCLIP ViT-L.

14. Movie information to prompt

This section introduces the process of converting movie information into a prompt. The attributes used include title, release_date, genres, keywords, and overview, with only the year extracted from the release_date using TMDB API. We input these attributes into Table 6.

ML prompt template for movie information using TMDB API
A movie poster of '{title}' from {release_date}, with genres including {genres}. Keywords are {keywords}, description is '{overview}'.

Table 6. Prompt template for movie information

15. Extracting key keywords

This section introduces the approach used to extract key keywords from Figures 1 and 4. We utilized ChatGPT 4o by inserting the user’s historical prompts into the ’{historical prompts}’ in Table 7. The recorded responses are shown in Table 8.

ChatGPT input template for extracting key keywords from historical user prompts
Your task is to extract key keywords from historical texts using no more than 10 words. List them in order, separated by commas. Please extract the keywords for the following historical text. The historical text: {historical prompts} The keywords:

Table 7. Prompt template for extracting key keywords

Figure	Dataset	User	Keywords
1	-	1	Grunge, Textured, Rough strokes, Vintage, Distressed
1	-	2	Ink wash, Sumi-e, Monochrome, Delicate lines, Traditional
1	-	3	Melancholic, Hand-drawn Textures, Dramatic Lighting, Purples and Blues
4	PIP	50396	Grove, Fox princess, Lycoris Radiata, Color film, Photography effect, Illustration, Ballet, Kitten, Egyptian costume, Greek dress
4	PIP	51447	Laura Palmber, Elegant art, Anime, Cinematic feel, Intricate details, Golden ratio, concept art, Tarot, Dragon princess, Trending in ArtStation
4	PIP	71263	Photorealism, Fantasy, Cuisine, Crystal, Goddess, Murals, Underground city, Halloween, Autumn, Future
4	ML	75	Mystery, Thriller, Horror, Journalist, War correspondent, Biography, Documentary, Politics, Revenge, Terrorism
4	ML	327	Love, Stand-up comedy, War, Submission, Jealousy, Murder, Avant-garde, Hostage, Alien, Crime
4	ML	410	Horror, Murder, Mystery, Silent film, Frontier, Gambler, Romance, Jazz, Terrorism, Kidnapping

Table 8. Key keywords of each user shown in Figures of the main paper.