Foundation Encoders Are All You Need for Preference-Aware Personalization

Supplementary Material

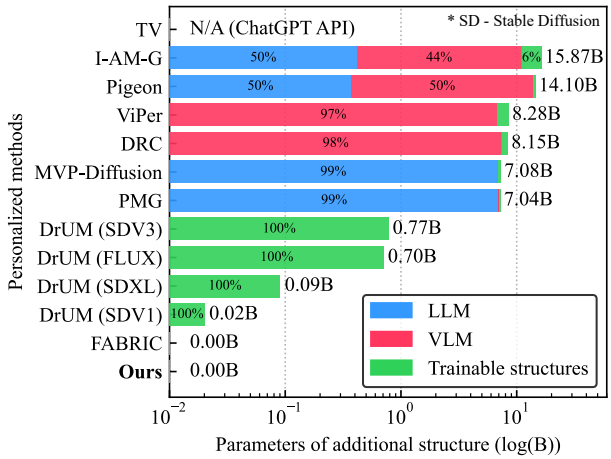


Figure 8. Parameter analysis for each personalized methods. Bars use a log scale, stacked colors indicate the proportions of each module, SD denotes Stable Diffusion.

CLIP score ↑		ML		
		Target	History	Imp
FLUX	-	22.42	13.25	-
	FABRIC	-	-	-
	TV	24.15	16.06	+14.46%
	PMG	-	-	-
	DrUM	25.12	16.16	+17.00%
	FAN	**26.16**	**16.27**	+19.71%

Table 4. Performance of FAN compared to baselines on PIP using FLUX. Imp denotes the average improvement rate over original models, and **bold** indicates the highest performance.

Rank priority ↑	FABRIC	TV	PMG	DrUM	Ours
PIP	-	1.77	-	1.87	**2.34**
ML	-	2.62	1.90	2.57	**2.92**

Table 5. Human evaluation based on ranking priority.

7. Parameter analysis of personalized methods

Figure 8 presents the resource requirements of user-behavior-based personalized image generation methods, excluding foundation text-to-image (T2I) models (*i.e.*, diffusion models, text encoders). I-AM-G and Pigeon adopt both LLM/VLM modules and additional trainable structures, demanding significant resources. ViPer, DRC, MVP-Diffusion, and PMG each rely on a single large-scale model with extra structures, exceeding 7B parameters. Although DrUM is relatively lightweight, its resource cost still grows with more modern architectures. FABRIC employs direct feature guidance without additional models but relies only on a single image pair on Stable Diffusion V1/V2, while TV requires API costs instead of local computation. In contrast, FAN achieves personalization without any extra resources.

8. Detailed comparison analysis

We provide additional comparative analysis with the baseline, which was not covered in Section 5.1. As shown in Table 1, FABRIC performs competitively on PIP in a resource-free manner but lacks generalization due to model-specific dependency. TV presents weak personalization even with ChatGPT 4o, performing worse than the original model in ML. PMG improves historical CLIP scores in ML but suffers from domain-specific dependency. Both TV and PMG depend on external large-scale models with limited references, compromising target fidelity.

9. More comparison on PIP using FLUX

Unlike OpenCLIP-based foundation T2I models (*e.g.*, Stable Diffusion V1/XL/V3), FLUX heavily relies on the Google T5 encoder. We evaluate CLIP scores on PIP using FLUX to verify FAN's generality, as shown in Table 4. FAN shows significant improvements on both target and history, demonstrating robustness across text encoder types.

10. Human evaluation

We invited 15 volunteers to rank 40 images generated from the PIP and ML datasets using Stable Diffusion V3 and FLUX. Participants were asked to prioritize the results based on target fidelity and user preference. As presented in Table 5, the average scores show that FAN achieved the highest performance among all baselines.

11. Detailed profiling comparison

Table 6 presents a detailed comparison of profiling methods, which was not covered in Section 5.1. TV_{CLIP} achieves high CLIP scores by selecting only similar entities, overlooking semantic diversity. TV_{BM25} focuses on keyword frequency using TF-IDF, failing to capture semantic relevance. PMG refers only to recent entities, leading to reduced overall performance. As discussed in the main paper, FAN achieves the best balanced performance, effectively capturing user preferences both semantically and structurally, while preserving target fidelity.

User profiling		PIP			ML		
		Target	History	Avg	Target	History	Avg
CLIP score ↑	Random	25.28	15.31	20.30	30.64	**12.83**	21.73
	Uniform	25.31	15.30	20.31	30.61	12.80	21.70
	TV_{BM25}	20.23	15.18	17.71	24.95	12.77	18.86
	TV_{CLIP}	**25.61**	14.87	20.24	**30.94**	12.75	**21.84**
	PMG	25.09	15.31	20.20	30.67	12.73	21.70
	Coreset	25.14	**15.51**	20.32	30.56	12.80	21.68
	Ours	25.47	15.24	**20.36**	30.69	**12.83**	21.76
Text align ↑	Random	94.50	60.09	77.30	96.87	38.54	67.71
	Uniform	94.81	60.16	77.48	96.88	38.58	67.73
	TV_{BM25}	92.56	58.44	75.50	96.77	38.41	67.59
	TV_{CLIP}	**96.63**	56.84	76.73	97.42	38.10	67.76
	PMG	92.31	58.75	75.53	96.64	38.41	67.67
	Coreset	92.13	**60.66**	76.39	96.87	**38.90**	**67.88**
	Ours	95.94	59.16	**77.55**	96.94	38.64	67.79

Table 6. Detailed performance of user profiling methods using Stable Diffusion V1 and OpenCLIP ViT-L. **Bold** and underline indicate the highest and second-highest performance.

User profiling		PIP			ML		
		Target	History	Avg	Target	History	Avg
CLIP score ↑	AVG	25.47	**15.24**	**20.36**	**30.69**	12.83	**21.76**
	CLS	**25.48**	15.16	20.32	30.65	**12.85**	21.75
Text align ↑	AVG	95.94	**59.16**	77.55	**97.00**	**38.81**	**67.91**
	CLS	**96.25**	58.91	**77.58**	96.95	38.66	67.81

Table 7. Performance of user profiling under AVG condition and CLS token in k-center greedy using Stable Diffusion V1 and OpenCLIP ViT-L. **Bold** indicates the highest performance.

User profiling		PIP			ML		
		Target	History	Avg	Target	History	Avg
CLIP score ↑	One-time	25.47	15.24	20.36	**30.69**	12.83	**21.76**
	Adaptive	**25.48**	**15.27**	**20.38**	30.59	**12.87**	21.73
Text align ↑	One-time	95.94	59.16	77.55	96.94	38.64	67.79
	Adaptive	**96.00**	**59.53**	**77.77**	**97.00**	**38.81**	**67.91**

Table 8. Performance of user profiling with one-time and adaptive strategies during the generation process using Stable Diffusion V1 and OpenCLIP ViT-L. **Bold** indicates the highest performance.

12. Enhancement of tailored profiling

Compared to prior methods [3, 11] that rely on the class (CLS) token in OpenCLIP text encoders, our tailored profiling remains effective even when using the averaged (AVG) condition during the k-center greedy process, as shown in Table 7. We further adopt a two-stage ranking strategy [5], proven efficient in retrieval research, to sample representative references from each foundation encoder. This allows adaptive diversity reflection when profiling is applied to each Personalized Attention (PA) layer, as shown in Table 8. Although adaptive profiling is highly effective, we adopt a one-time profiling for practical efficiency in real-world scenarios.

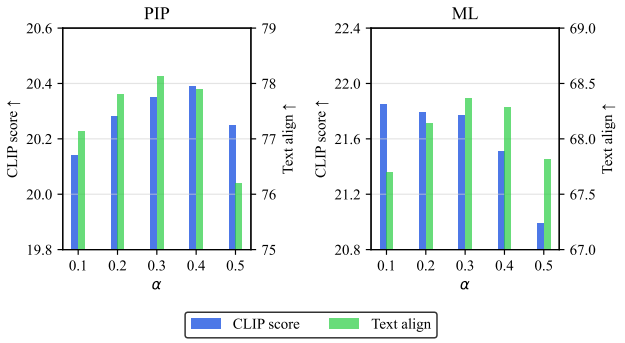


Figure 9. Average performance between target and history for different personalization degree α using Stable Diffusion V1 and OpenCLIP ViT-L.

13. Impact of personalization degree

Figure 9 shows the average performance under different personalization degrees α. In PIP, FAN achieves its best scores around 0.3–0.4, and in ML between 0.2–0.3. When α exceeds these ranges, target fidelity drops due to over-personalization. Therefore, we set α to the optimal value that maximizes average performance.

14. Personalized degree in target fidelity

Figure 6 in the main paper illustrates how visual attributes change with varying personalization degrees α. Additionally, Figure 10 evaluates whether reference attributes are faithfully incorporated while robustly maintaining target fidelity. For instance, in (a), the painter is smoothly blended with abstract concepts (*e.g.*, old age, transcendence) as personalization increases, while in (b), the astronaut is naturally incorporated into the natural scenery (*e.g.*, space, snowy mountain) and overall atmosphere (*e.g.*, solitude, retro-futuristic, color).

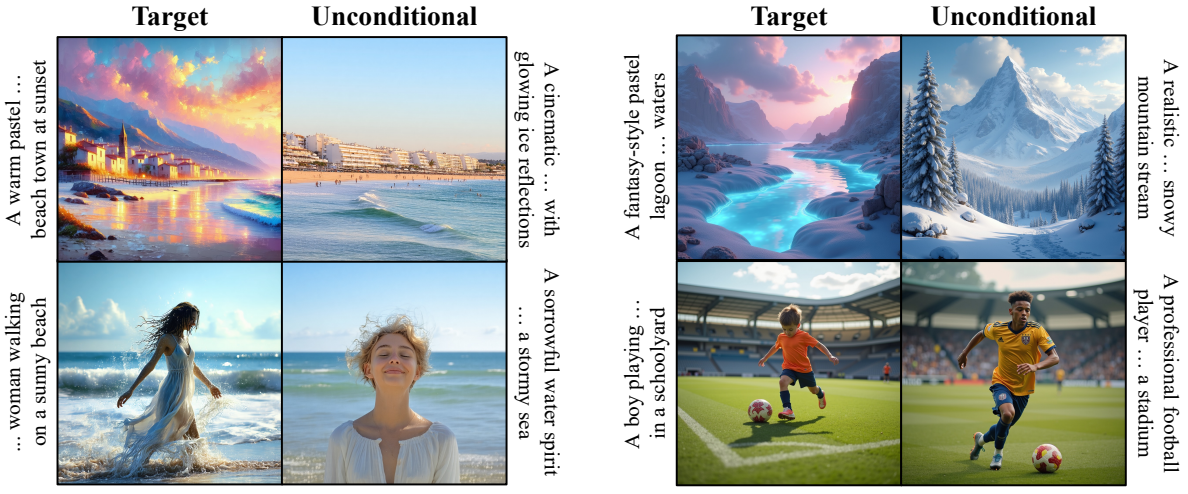
15. Impact of personalized query

Figure 11 visualizes performance differences between target and unconditional queries for personalized inputs. Unconditional queries cause semantic loss during personalization, whereas target queries preserve semantic expressiveness and incorporate reference attributes effectively. For instance, in (a), unconditional queries lose contextual details (*e.g.*, pastel, sunset, walking, sorrowful, stormy), while in (b), unconditional ones fail to preserve the scene semantics (*e.g.*, pastel, lagoon, water, boy).

(a) Stable Diffusion V3

(b) FLUX

Figure 10. Impact of personalization degree α between single target and dual references using Stable Diffusion V3 and FLUX.



(a) Stable Diffusion V3

(b) FLUX

Figure 11. Effect of using target or unconditional queries as the personalized query. Left and right texts correspond to the target and reference prompts, respectively.

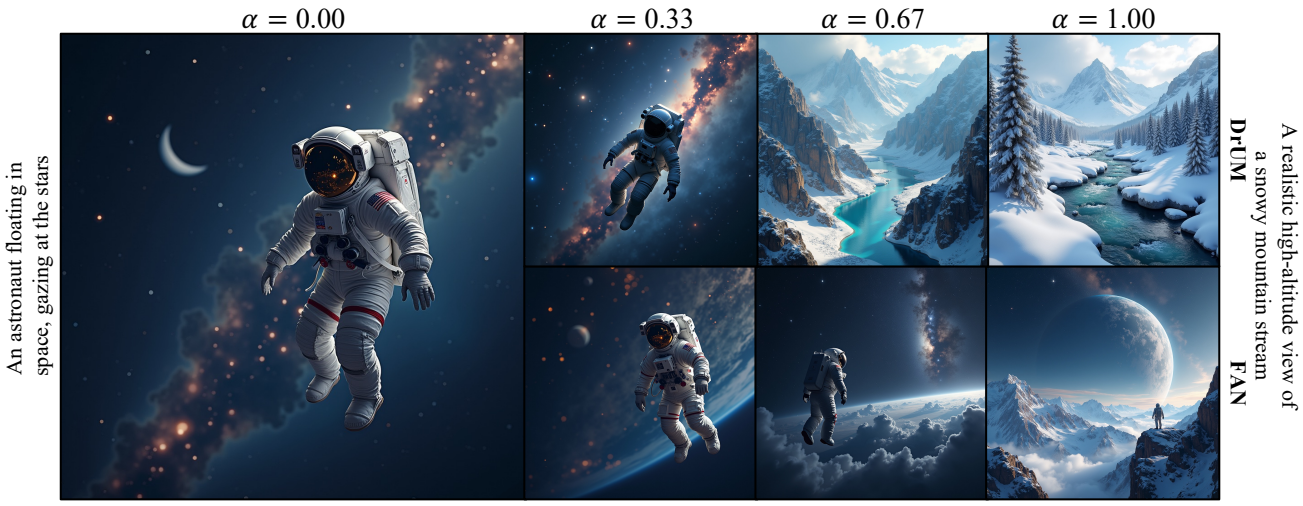


Figure 12. Personalization difference between DrUM and FAN.

User 5 Sepia illustration, vintage tones, soft shading, aged texture, nostalgic mood

Figure 13. Comparison of personalization across foundation models using Stable Diffusion V3 and FLUX.

16. Target fidelity on personalization

Figure 12 shows the difference in personalization on FLUX. Compared to DrUM, FAN maintains target fidelity across personalization degrees α, even under substantial contextual shifts.

17. Model characteristics on personalization

As shown in Figures 1 and 5 of the main paper, Figure 13 highlights that FAN preserves each model's unique creativity and diversity even for the same user. Since FAN reconstructs foundation encoders without additional training or personalized structures, it achieves personalization while maintaining characteristics inherent to each architecture.

18. Beyond individual-level personalization.

Figure 14 shows the extensibility to new scenarios (*i.e.* groups and brands). Rather than using the target query, simply replacing the personalized query with a group or brand prompt allows integrating such broader context within the personalized conditioning. This suggests that the proposed approach can be naturally extended to new tasks and applications beyond individual-level personalization.

Figure 14. Synthesis results of group- and brand-based conditioning. Left and right texts denote target and reference prompts, respectively.

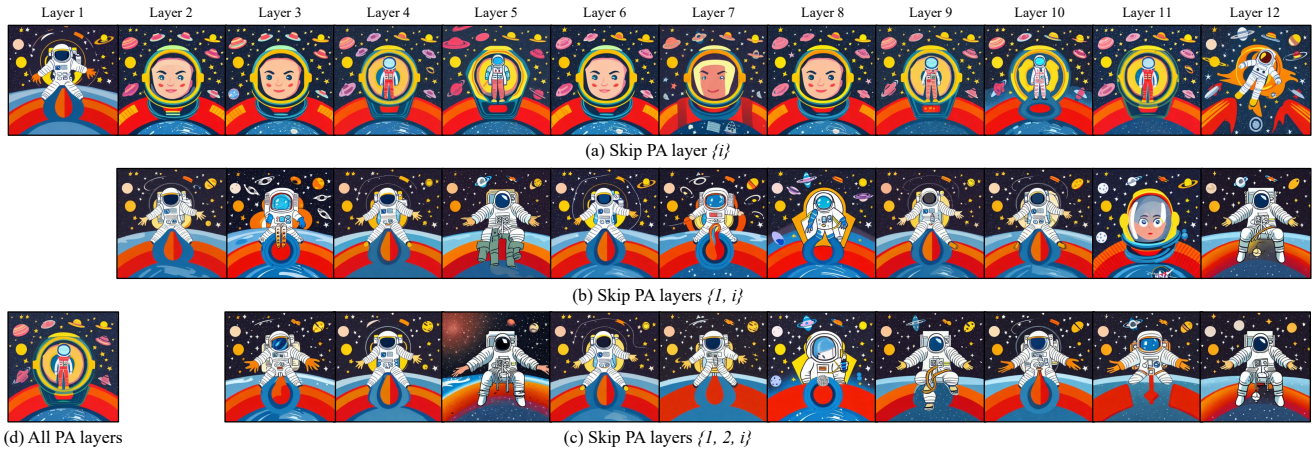


Figure 15. Impact of the PA trade-off using Stable Diffusion V1. Target and reference prompts used are: "An astronaut floating in space, gazing at the stars." and "A retro-futuristic space exploration movie poster with bold, vibrant colors.".

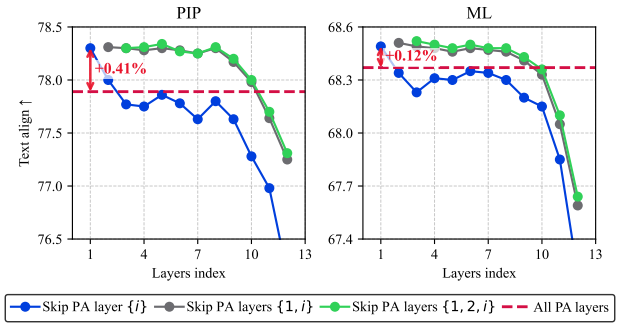


Figure 16. Impact of PA trade-off using OpenCLIP ViT-L.

19. PA trade-off analysis.

Figure 16 presents the effect of applying PA, showing the average text alignment (text align) over target and history. The red line denotes performance when all layers are replaced with PA. Skipping PA only in the first layer yields the highest score by refining target representation. Excluding PA in a few early layers offers marginal gains, but such effects depend on datasets and reduce generality. Figure 15 further visualizes this trend in Stable Diffusion V1, confirming that skipping PA only in the first layer enhances target fidelity while maintaining stable personalization. Therefore, FAN skips PA solely in the first layer, achieving a balanced trade-off without additional optimization.

20. Detailed information

This section introduces prompts, keywords, and details used in our study. To convert movie descriptions and extract key keywords, we employ prompt templates for ChatGPT 4o (see Tables 9 and 10), as introduced in [11]. The keywords and prompts used in the figures are described in Table 11.

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

Table 9. Prompt template for movie description

Prompt template for extracting key keywords
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 10. Prompt template for extracting key keywords

Figure	User or group	Keywords or prompts
1, 5	1	Watercolor painting, paper texture, colorful, artistic splashes
1, 5	2	Minimal line drawing, fine ink lines, soft shadows, ample negative space
1, 5	3	Digital art, neon lighting, futuristic glow, high contrast, vibrant palette
5	4	Flat illustration, geometric layout, clean lines, bright palette, minimal shading
5, 13	5	Sepia illustration, vintage tones, soft shading, aged texture, nostalgic mood
4	45894	Boy, Love tree, Flowers bloom, Antlers, Corgi, Decaying city, Future crisis, Grove, Starlight, Pipa, Moonlight
4	51268	Art nouveau, Alphonse mucha, Digital painting, Chinese palace, Gold dragon, Burning phoenix, Aurora, Fantasy library, Victorian window, Jewelry design
4	2945422	Beautiful Girl, French Bangs, Cartoon Style, Oval face, SpongeBob, High school student, Slim, Sunny, School uniform, Headphones
4	75	Mystery, Thriller, Horror, Journalist, War correspondent, Biography, Documentary, Politics, Revenge, Terrorism
4	127	Drama, Comedy, Crime, Detective, Ballet, Mental asylum, Stand-up comedy, Wrestling, Political scandal, Romance
4	410	Horror, Murder, Mystery, Silent film, Frontier, Gambler, Romance, Jazz, Terrorism, Kidnapping
6	-	A scholar floating in an infinite library of glowing knowledge orbs
6, 10, 14	-	A retro-futuristic space exploration movie poster with bold, vibrant colors
10	-	A grandmother knitting a sweater while watching TV in a living room
10	-	A cosmic dreamweaver crafting galaxies from swirling nebulae
10, 11	-	A realistic high-altitude view of a snowy mountain stream
11	-	A warm pastel painting of a beach town at sunset
11	-	A cinematic winter valley with glowing ice reflections
11	-	A cheerful young woman walking on a sunny beach
11	-	A sorrowful water spirit dissolving into a stormy sea
11	-	A fantasy-style pastel lagoon with glowing waters
11	-	A boy playing with a soccer ball in a schoolyard
11	-	A professional football player training in a stadium
14	-	A spirit guardian in the form of a spectral wolf, watching over the spirit realm

Table 11. Keywords and prompts shown in Figures of the main paper.