ITERDIFF: TRAINING-FREE ITERATIVE FACE EDITING VIA EFFICIENT CLIP-GUIDED MEMORY BANK
SUPPLEMENTARY MATERIALS

1. ITEREDITBENCH

To generate the instruction set (Tabel 1), we use the following prompt:

"Create an instruction set for facial image editing tasks. Each category (e.g., age, gender, skin tone, hair) should have a template and associated attributes. Use the following format:
"'python
INSTRUCTIONS = {
"<Template>": [
"<Attribute 1>",
"<Attribute 2>",
],
}
"'
Examples include:
- Template: "Make the face look {}."
Attributes: ["older", "like a teenager", "middle-aged"]
- Template: "Change the gender to {}."
Attributes: ["male", "female"]
Generate templates and attributes that ensuring diversity and clarity."

Table 1: Instruction set.

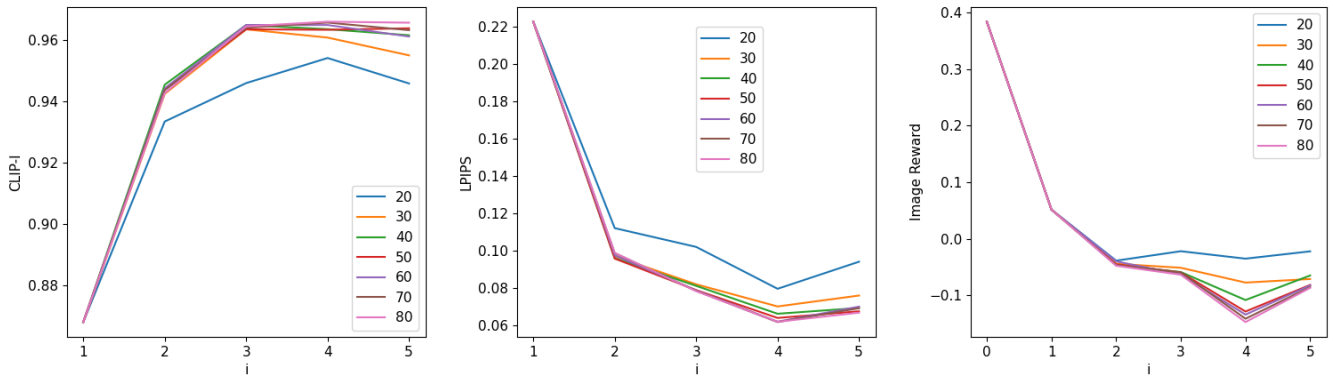
Template	Attributes
"Make the face look {}."	*"older"*
	"younger"
	"more mature"
	"childlike"
"Change the gender to {}."	*"male"*
	"female"
"Make the person's skin {}."	*"darker"*
	"lighter"
	"paler"
	"tanned"
"Give {} appearance."	*"an Asian"*
	"an Indian"
	"a Middle Eastern"
	"a Western European"
	"an African"
"Change the hair color to {}."	*"brown"*
	"blonde"
	"black"
	"red"
	"gray"
"Make the hair {}."	*"shorter"*
	"longer"
	"curlier"
	"straighter"
"Put on {}."	*"glasses"*
	"sunglasses"
	"a cap"
	"a scarf"
"Add {} to the face."	*"a beard"*
	"a mustache"
	"a light stubble"
	"a goatee"

2. ABLATION STUDY

2.1. Impact of the Applied Range in *ECMB* (s)

As shown in Fig. 1(a), CLIP-I improves rapidly with increasing s and stabilizes around $s = 40$, indicating enhanced semantic consistency. Similarly, LPIPS (Fig. 1(b)) decreases as s increases, suggesting better perceptual similarity, but beyond $s = 40$, the improvement becomes marginal. Regarding image quality, ImageReward (Fig. 1(c)) follows a similar trend, where overly large s values slightly degrade realism due to excessive constraints. Based on these observations, we select $s = 40$ as the optimal value, balancing content preservation and high-quality edits in iterative face editing.

Based on these observations, we select $s = 40$ as the optimal value, balancing content preservation, instruction alignment, and high-quality edits in iterative face editing.



(a) (b) (c)

Fig. 1: Quantitative curves for different s, k is fixed to 20 here.

2.2. Effect of Selecting Different Numbers of Pairs (k)

We examine the influence of k on iterative editing performance by analyzing metrics such as CLIP-I, LPIPS, and ImageReward across different values of k (see Fig. 2). As depicted in Fig. 2(a), CLIP-I improves notably as k increases, plateauing around $k = 20$, which indicates stable semantic consistency. Similarly, LPIPS scores, shown in Fig. 2(b), decrease as k increases, suggesting enhanced perceptual similarity. However, the gains beyond $k = 20$ become negligible. For image quality, as demonstrated in Fig. 2(c), ImageReward shows a comparable pattern, where excessive values of k negatively impact realism. This suggests that retaining too much prior information can hinder the adaptation to new instructions, reducing the effectiveness of edits. Taking these findings into account, we adopt $k = 20$ as the optimal setting, achieving a balance between content fidelity, adherence to instructions, and high-quality iterative edits.

2.3. Guidance Factor (g^i)

To evaluate the effectiveness of the Guidance Factor, we conduct an ablation study comparing results with and without it. From Fig. 3(a) and (b), we observe that removing the Guidance Factor leads to higher CLIP-I and lower LPIPS, indicating better alignment with the original image and higher perceptual similarity. However, this improvement comes at the cost of reduced edit strength, making the modifications less distinguishable. In contrast, incorporating the Guidance Factor achieves a better balance by enabling more pronounced edits.

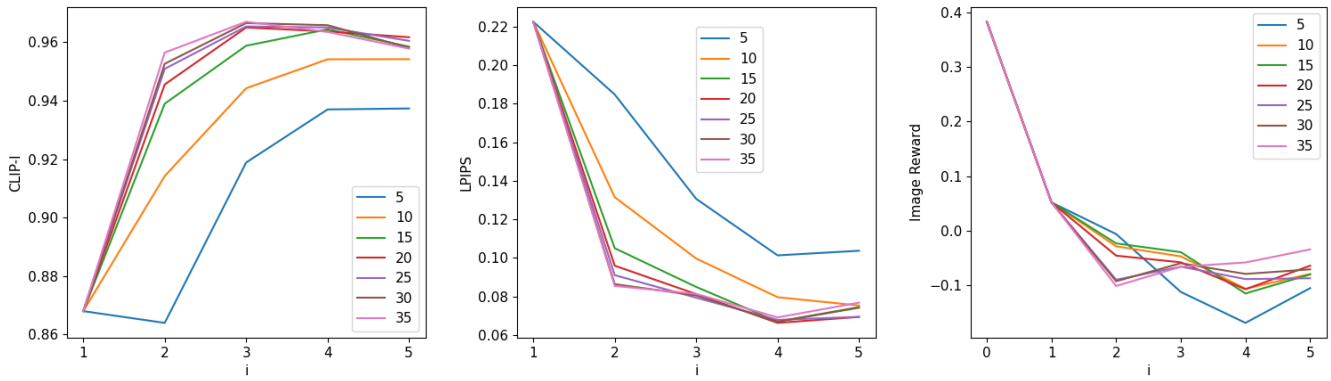
Additionally, from Fig. 3(c), we see that the ImageReward metric is improved with the Guidance Factor, supporting the claim that our method generates images that better match the given textual instruction and enhances edit quality while maintaining semantic consistency.

Overall, this ablation study highlights the importance of the Guidance Factor in achieving meaningful and effective image editing.

2.4. TF^{2}P

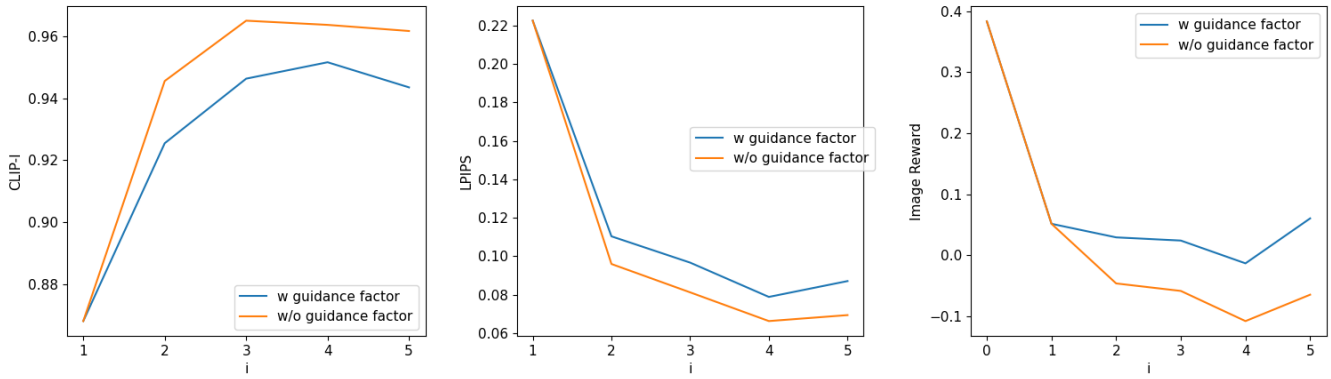
To further investigate whether TF^{2}P retains redundant information, we conduct an additional experiment setting $s = 100$ and $k = 100$, saving all KV pairs and using all KV pairs from the previous edit. From Fig. 4, we observe that TF^{2}P consistently achieves higher CLIP-I scores compared to ECMB (Fig. 4(a)), indicating that it preserves stronger image similarity with the input. Additionally, TF^{2}P achieves lower LPIPS scores (Fig. 4(b)), further confirming that the edited images maintain a high level of perceptual similarity. However, while these metrics suggest better preservation, they also imply that the edits may not be as pronounced, potentially leading to insufficient modification strength. On the other hand, in terms of ImageReward (Fig. 4(c)), ECMB consistently outperforms TF^{2}P, indicating that it better aligns with the textual prompt and reinforcing its ability to generate images that are more aesthetically pleasing and contextually relevant.

These results indicate that TF^{2}P retains too much information from previous edits, which hinders the clarity and effectiveness of modifications, making the editing effect less distinct.



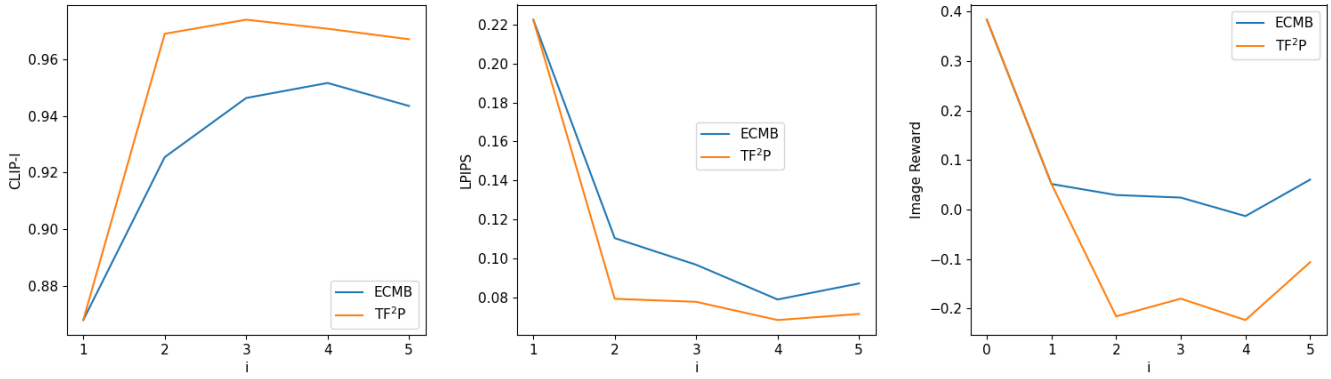
(a) (b) (c)

Fig. 2: Quantitative curves for different k, s is fixed to 40 here.



(a) (b) (c)

Fig. 3: Quantitative curves for with or without guidance factor.



(a) (b) (c)

Fig. 4: Quantitative curves for ECMB and TF^{2}P.