Gen-AFFECT: Generation of Avatar Fine-grained Facial Expressions with Consistent identiTy Supplementary Material

Hao Yu
Boston University
Boston, MA

Rupayan Mallick
Georgetown University
Washington, D.C.

Margrit Betke
Boston University
Boston, MA

Sarah Adel Bargal
Georgetown University
Washington, D.C.

A. Ethical Impact Statement

Our work focuses on avatar generation with fine-grained facial expressions, aiming to support applications in education, gaming, and virtual communication. While such technologies have the potential to enrich user engagement and experience, they also raise important ethical considerations. As with any generative model, our framework carries risks of misuse. These include the potential to generate misleading or harmful content, impersonate individuals, replicate creative work without proper credit, and compromise personal privacy. Additionally, the ability to manipulate expressions may enable deceptive emotional cues or impersonation in virtual contexts.

We acknowledge these concerns and are committed to promoting the responsible development and use of generative technologies. This includes ensuring transparency, promoting ethical deployment, and explicitly discouraging the use of our model for deceptive, malicious, or privacy-infringing purposes.

B. Ablation Study

We conducted ablation studies evaluating the effects of removing the consistent attention module and removing the expression text in the input prompt (Table 1 and Figure 2). We observe that removing the consistent attention module leads to a significant drop in image consistency. From the qualitative results in Figure 2, the generated avatars exhibit inconsistencies in appearance, such as varying hair colors and noticeable differences in facial features. Removing expression text from the prompt results in less expressive generations and lower expression-related metrics. Note that the improvement in consistency scores is due to the outputs becoming more neutral and less expressive, making them visually more similar rather than better controlled. Lastly, while we ablated the components above to analyze their individual impact, we chose not to ablate the three proposed loss terms individually as they are jointly essential for generating expression- and identity-controlled avatar images. The flow loss serves as the primary training objective for image generation, while the identity and expression loss ensure that the generated image reflects the intended identity and expression. Removing any of these losses results in a failure in generation. Without flow loss, the model cannot generate reasonable images. Without identity or expression loss, the model degrades into a generic T2I model with no control over identity or expression. Given this, ablating these losses would not yield meaningful insights.

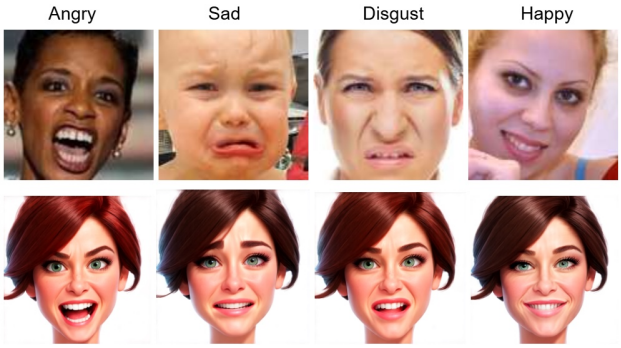

Figure 1. Generated avatar using expression images from AffectNet.

C. Out-of-Distribution Expression Input

In this work, we use expression images from Emo135 dataset [1] as our reference expression images. While expression images from the same dataset may yield the best performance, our model supports arbitrary expression images at inference time. To demonstrate this generalization, we tested our model with images from AffectNet [2], a dataset of basic facial expressions, and found that it can still successfully generate the corresponding expressions (Figure 1).

D. Expression embeddings on Emo135 dataset.

Other than POSTER expression embeddings [3], we additionally trained expression embeddings on the Emo135 dataset [1], which provides complementary insights into

	Expression		Identity		Consistency	
Model	Exp.↓	CLIP↑	ID.↑	DINO↑	DINO Con.↑	ID Con.↑
Full model	**11.09**	**0.678**	0.361	**0.828**	0.957	0.762
w/o consistent attention	11.47	0.683	0.361	0.816	0.934	0.723
w/o expression prompt	12.25	0.643	**0.363**	0.815	**0.981**	**0.889**

Table 1. Ablation study showing the effect of removing consistent attention and expression prompt.

Table 2. Results of expression errors using expression embeddings trained on the Emo135 dataset.

Model	Emo. Exp.↓
FastComposer	30.08
PuLID	29.86
PhotoMaker	29.52
Conditional SDXL	28.89
Gen-AFFECT (Ours)	**28.84**

our model's ability to capture fine-grained expressions. We evaluated our model using these new embeddings and report the results in Table 2. Our model consistently outperforms previous methods.

References

[1] Keyu Chen, Xu Yang, Changjie Fan, Wei Zhang, and Yu Ding. Semantic-rich facial emotional expression recognition. *IEEE Transactions on Affective Computing*, 13(4):1906–1916, 2022. 1

[2] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 1

[3] Ce Zheng, Matias Mendieta, and Chen Chen. POSTER: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 3146–3155, 2023. 1

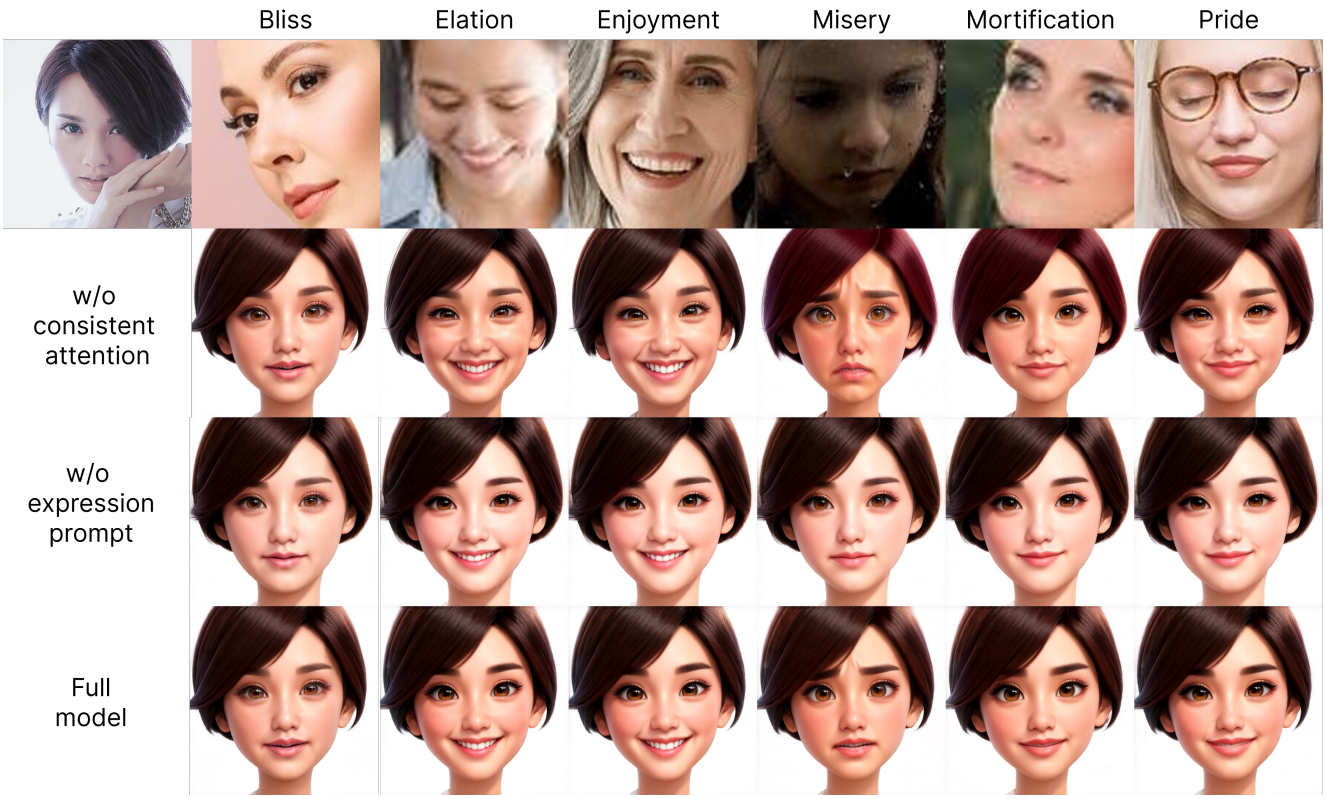

Figure 2. Qualitative results for ablation study. After removing the consistent attention module, the generated avatars exhibit inconsistencies in appearance, such as varying hair colors and noticeable differences in facial features. Removing expression text from the prompt results in less expressive generations.