

Supplementary Materials: TextGaze: Gaze-Controllable Face Generation with Natural Language

Anonymous Authors

1 DIVERSITY OF EXTENT WORDS

LLMs are famous for the great diversity of generated languages. Our dataset generated by LLMs inherits this advantage. Our ToG dataset has 475 different adverbs for extent description. In contrast, the well-known text pose datasets BABEL [4] and PoseScript [1] only have 250 unique action categories and 87 unique posecodes separately. We show some examples of extent description words in Table. 1.

2 CONTROLLING SHAPE AND EXPRESSION

Thanks to our sketch-guided model and the 3D face model, we can control the face shape and expression by changing the input sketches by interpreting the shape and expression code of the 3D face model. We show the visualization results in Fig. 1. It shows that our model can generate faces with different shapes and expressions while keeping the head poses and gaze directions the same, which enriches the potential applications of our model.

3 ADDITIONAL QUALITATIVE RESULTS

The additional results of our model are presented in Fig. 2. It contains 12 sets of results including text description and generated image. They are in three different styles. For better visualization, We annotate the possible gaze directions using the green sector and the head direction with a blue arrow same as the main paper. Please note the blue arrow only shows left/right direction. Please refer to the text for head pose in the vertical direction. The close alignment of the text descriptions with the generated images demonstrates the effectiveness of our model.

4 BUDGET EFFICIENCY OF THE GENERATED DATA

We compare the budget efficiency against the manual annotation on Amazon Mechanical Turk. According to the academic report [2], Amazon Mechanical Turk pays between 1 dollar and 6 dollars per hour for a typical Human Intelligence Task (HIT) task. In the comparison, we use the lowest price which is 1 dollar per hour. We assume each worker can annotate one pair of head pose and gaze direction with two text descriptions in one minute. For LLMs annotation, we use ChatGPT4 Turbo. According to [3], 1 token is approximately 0.75 words. The price is 0.01 dollars per 1K tokens for input and 0.03 dollars per 1K tokens for output.

Based on the above price, we show the estimated cost of our annotation task using Amazon Mechanical Turk and ChatGPT4 Turbo in Table 2. It indicates that using LLMs like ChatGPT has better budget efficiency than the manual annotation on Amazon Mechanical Turk.

REFERENCES

- [1] Delmas, Ginger and Weinzaepfel, Philippe and Lucas, Thomas and Moreno-Noguer, Francesc and Rogez, Grégory. 2022. PoseScript: 3D Human Poses from Natural

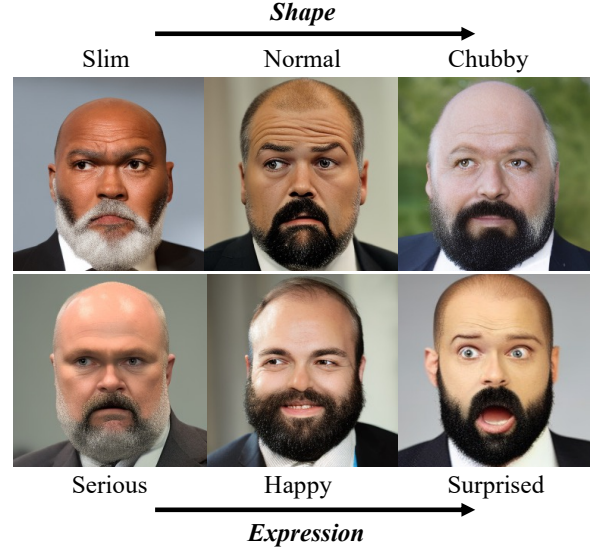


Figure 1: Visualization of interpreting shape and expression.

Table 1: Examples of extent synonym in ToG.

Extent	Synonym Examples in ToG
Slightly	mildly, faintly, subtly, marginally, a little, a bit, gently, minimally, ...
Moderately	reasonably, medium, fairly, modestly, somewhat, gradually, plainly, gracefully, ...
Significantly	considerably, substantially, notably, sharply, seriously, majorly, vastly, heavily, ...

Table 2: Comparison on costs.

Annotation Method	Costs (\$) ↓
Amazon Mechanical Turk	796.23
ChatGPT4 Turbo	114.02

Language. In ECCV.

- [2] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.
- [3] OpenAI. 2024. Pricing. <https://openai.com/pricing>.
- [4] Abhinanda R Punnakal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. 2021. BABEL: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 722–731.



Figure 2: Additional qualitative results. The blue arrow only shows left/right direction. Please refer to the text for head pose in the vertical direction.