Review of: "Identity-Preserving Text-to-Video Generation by Frequency Decomposition"

Tao Lu[1]

1 Independent researcher

Potential competing interests: No potential competing interests to declare.

1. This is my first time reviewing on this platform, and I am not sure if my evaluation is appropriate. Are the low-frequency and high-frequency features mentioned by the author in the manuscript, "low-frequency" and "high-frequency" relative concepts? Or do they have specific representative meanings?
2. The dynamic cross-face loss mentioned by the author uses Gaussian noise to enhance generalization ability and prevent copy-paste generation. How should the intensity of Gaussian noise be set here? When the signal-to-noise ratio is still high after adding Gaussian noise, can it effectively prevent copy-paste generation?
3. I really want to raise the question of whether the evaluation index is consistent with human perception. The author has put this part in future work. If possible, it is recommended that the author make appropriate supplements in this part, which will make the content more substantial.
4. The article mentions the importance of high-frequency and low-frequency features for identity preservation, but are there more theories or experiments to support the rationality and effectiveness of this decomposition?
5. Is the current method limited to face generation? If applied to scenes containing multiple people and multiple objects, what is the scalability and performance of the model?
6. Is there any bias in the training dataset, such as whether the gender, age, or race distribution is balanced? Will these biases affect the identity preservation effect of the generated video?
7. High-frequency features are fused through CLIP and the facial recognition backbone network, but the article does not mention the impact of noise or bias on the extraction of these features. Is there any further verification?
8. How are the hyperparameters of dynamic mask loss and dynamic cross-face loss determined? Are there any experiments to verify the sensitivity of these parameters?
9. Is there any theoretical support for the choice of injecting frequency signals into the Transformer attention block? Have other possible injection locations been explored?