Qeios

Peer Review

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

Yu Lingyun[1]

1. Independent researcher

This paper proposes an identity-preserving text-to-video generation framework based on a Diffusion Transformer architecture. With a frequency-aware heuristic identity-preserving control scheme, it achieves a tuning-free pipeline without case-by-case finetuning. Meanwhile, a hierarchical training strategy is proposed to facilitate training and enhance generalization. Extensive experiments demonstrate the effectiveness of ConsisID in generating high-quality, editable, and consistent identity-preserving videos under the frequency-aware DiT-based control scheme.

However, there still exist some questions that need to be solved:

1. Since the authors claim that there's greater difficulty for DiT-based models in training convergence and a weakness in perceiving facial details, the reason why the DiT-based pipeline should be used to implement the identity-preserving text-to-video task hasn't been presented.
2. The visualization of low-frequency and high-frequency tokens is presented in Figure 2, yet the method section does not include any introduction to the corresponding visualization process. Why does the output of the Q-Former first point to the visualization, which then points to the tokens? Additionally, the symbols of Equations (2) and (4) do not correspond accurately to Figure 2.
3. The definition and delineation of high and low frequency information is not clear, e.g., why should the semantic information extracted by the CLIP model belong to high frequency in Section 3.2.2?
4. As shown in Figure 2, three layers of features from the face recognition backbone are used to represent the intrinsic face identity. More ablation experiments should be given to incorporate

 doi.org/10.32388/O51IWN

more or fewer layers and verify if the current experimental setup introducing the three-layer features is optimal.

5. The amount of data and GPU resources used for model training are not provided. For a fair comparison, the authors should provide comparative experiments with SOTAs concerning model efficiency, such as inference time, model parameters, etc.

Declarations

Potential competing interests: No potential competing interests to declare.