Generative Video Diffusion for Unseen Novel Semantic Video Moment Retrieval

Dezhao Luo[1], Shaogang Gong[1], Jiabo Huang[2], Hailin Jin[3], Yang Liu[4,5]
[1]Queen Mary University of London, [2]Sony AI, [3]Adobe Research,
[4]WICT, Peking University, [5]State Key Laboratory of General Artificial Intelligence, Peking University,

SCAN ME

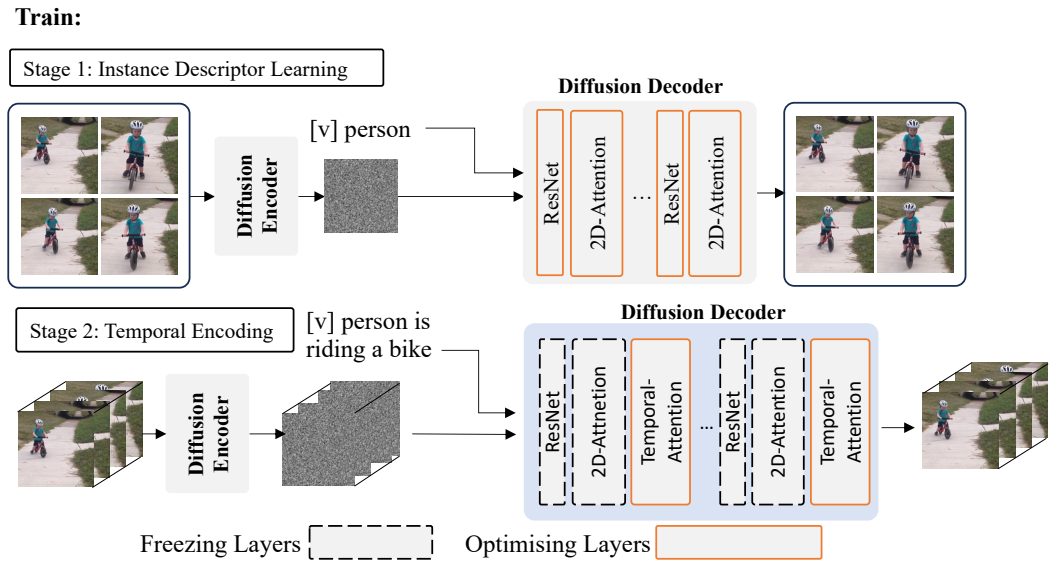
1. Problem Definition

The correlation between video moments and text is crucial for the task of **video moment retrieval (VMR)**, yet there is a scarcity of large-scale datasets.

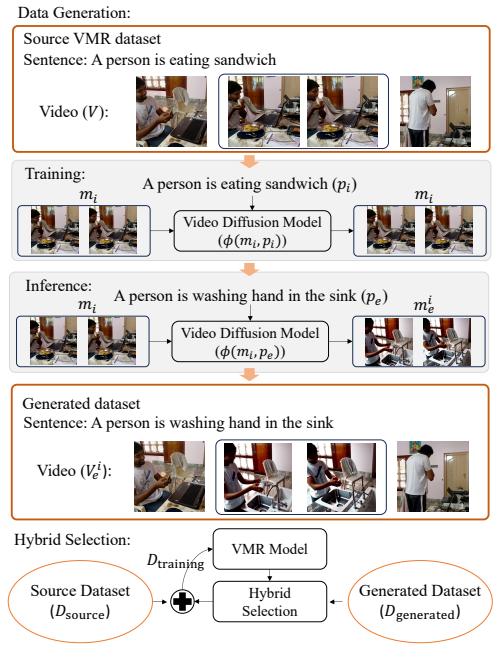
2. Solution

- A video diffusion model that synthesises training data
- A data selection module that selects beneficial data for the VMR task

3. Video Diffusion Model



4. Data Generation



5. Data Selection

Cross-modal relevance: $$s_c(p_e, m_e) = \frac{1}{N}\sum_{i=1}^{N} \cos(\text{VLM}(p_e), \text{VLM}(f_{m_e}^i))$$

Uni-modal structure: $$s_u(m_s, m_e) = \frac{1}{N}\sum_{i=1}^{N} \cos(\text{VM}(f_{m_s}^i), \text{VM}(f_{m_e}^i))$$

Model performance: $$D_{\text{mpd}} = \text{TOP}_l(\{(d, -\text{VMR}(d)) \mid d \in D_{cu}\})$$

6. Video Editing Ability

7. Conclusion

a. FVE changes the action in a video and maintains other details.
b. FVE generates high-quality training data that benefits the VMR task (44.89% vs. 44.01%).