

Generative Video Diffusion for Unseen Novel Semantic Video Moment Retrieval

Dezhao Luo¹, Shaogang Gong¹, Jiabo Huang², Hailin Jin³, Yang Liu^{4,5}

¹Queen Mary University of London, ²Sony AI, ³Adobe Research,

⁴WICT, Peking University, ⁵State Key Laboratory of General Artificial Intelligence, Peking University,

Problem Definition

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

Solution

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

Data Generation

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

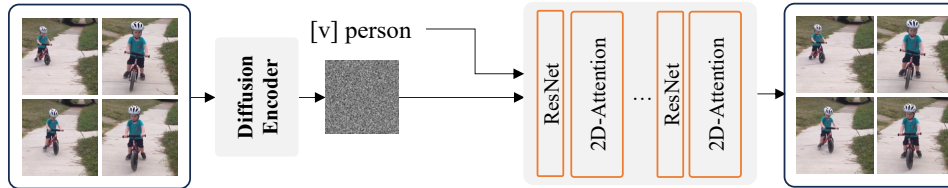
Uni-modal structure: $s_u(m_s, m_e) = \frac{1}{N} \sum_{i=1}^N \cos(\text{VM}(f_{m_s}^i), \text{VM}(f_{m_e}^i))$

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

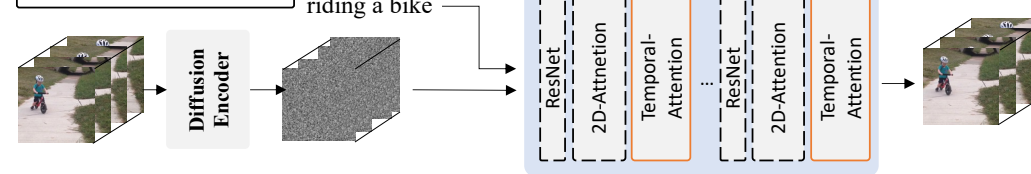
Video Diffusion Model

Train:

Stage 1: Instance Descriptor Learning



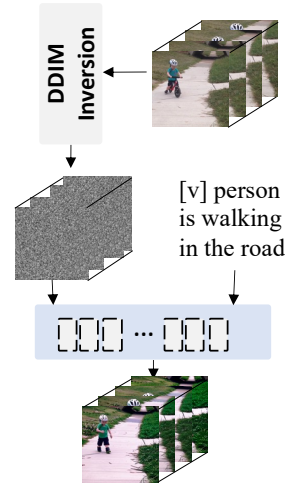
Stage 2: Temporal Encoding



Freezing Layers

Optimising Layers

Inference:

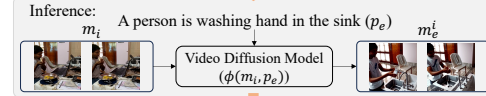
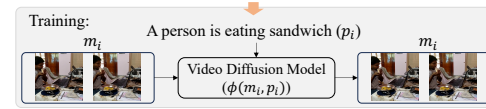


Data Generation

Data Generation:

Source VMR dataset

Sentence: A person is eating sandwich

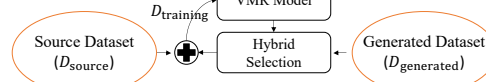


Generated dataset

Sentence: A person is washing hand in the sink



Hybrid Selection:



Video Editing Ability



Conclusion

- FVE is able to generate training data that benefits the VMR task
- FVE is able to change the action in a video and maintain other details.