

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

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

5. Data Selection

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

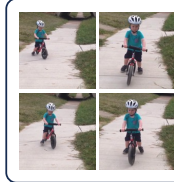
Uni-modal structure: $s_u(m_s, m_e) = \frac{1}{N} \sum_{i=1}^N \cos(\text{VM}(f_{m_s}^i), \text{VM}(f_{m_e}^i))$

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

3. Video Diffusion Model

Train:

Stage 1: Instance Descriptor Learning



[v] person

Diffusion Decoder

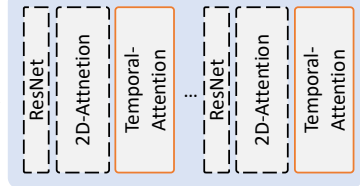


Stage 2: Temporal Encoding



[v] person is riding a bike

Diffusion Decoder



Freezing Layers

Optimising Layers

Inference:



[v] person is walking in the road



6. Video Editing Ability



4. Data Generation

Data Generation:

Source VMR dataset

Sentence: A person is eating sandwich

Video (V):

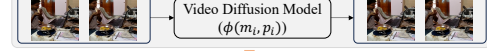


Training:

m_i

A person is eating sandwich (p_i)

m_i

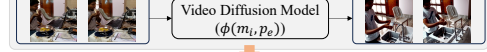


Inference:

m_i

A person is washing hand in the sink (p_e)

m_e



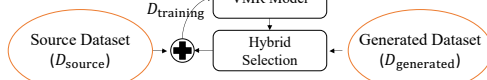
Generated dataset

Sentence: A person is washing hand in the sink

Video (V_e^i):



Hybrid Selection:



7. Conclusion

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