# 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,

## 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

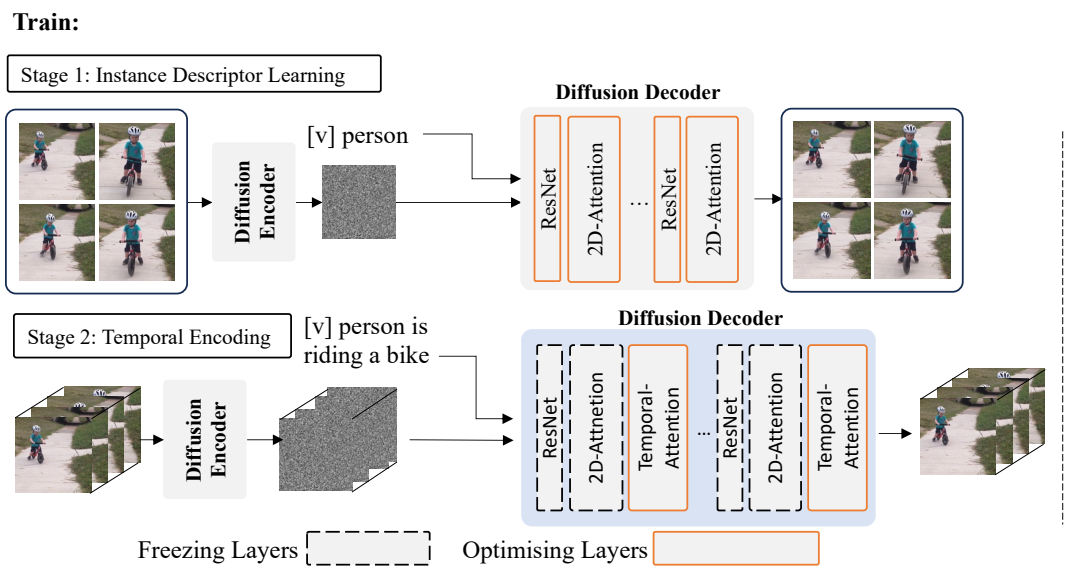
## Data Generation

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

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

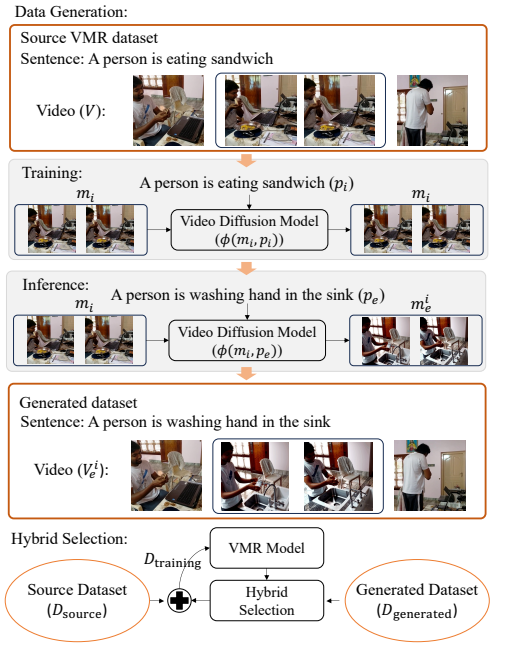
Model performance: $D_{\text{mpd}} = \text{TOP}_l(\{(d, -\text{VMR}(d)) \mid d \in D_{cu}\})$

## Video Diffusion Model




## Video Editing Ability


## Data Generation




## 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.