PVChat: Personalized Video Chat with One-Shot Learning

Yufei Shi[1*] Weilong Yan[2*] Gang Xu[3] Yumeng Li[4] Yucheng Chen[1] Zhenxi Li[1]
Fei Yu[3] Ming Li[3†] Si Yong Yeo[1†]

[1]Lee Kong Chian School of Medicine, NTU Singapore [2]National University of Singapore
[3]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) [4]Nankai University
* Equal Contribution † Corresponding author

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Code: https://github.com/DavidYan2001/PVChat
Email: yufei005@e.ntu.edu.sg

Problem Definition and Contributions

Goal: Enable one-shot, identity-aware video understanding and QA from a single reference video, generalizing across scenes and multi-subject settings

Motivation:

- General-purpose ViLLMs lack reliable person-specific reasoning due to non-personalized training.
- Prior personalization is image-centric, missing motion/temporal cues essential for video.

Key Contributions:

- PVChat: First ViLLM to personalize from a single reference video, enabling one-shot, identity-aware video understanding and user-specific QA.
- Data pipeline: Single-clip augmentation and QA generation using off-the-shelf tools to create identity-preserving positives and look-alike hard negatives; we release a diverse dataset (6 scenarios, 304 source videos, 2,304 generated videos, 30k+ QA pairs).
- Modeling: ReLU-Routing Mixture-of-Heads plus Smooth Proximity Regularization and Head Activation Enhancement for stable training and stronger extraction of individual-specific cues.

SPA and HAE Loss

SPA (Smooth Proximity Regularization)

- To encourage the sparsity of activation, which helps to reduce computational cost and boost specific-domain learning. T_s and R_s is the Target and current sparsity.

$$\mathcal{L}_{SPR} = \beta_p \cdot \mathcal{L}_{Reg} \qquad \mathcal{L}_{Reg} = \left\| \frac{1}{n}(\boldsymbol{W_r x_t}) \right\|$$

$$\beta_{p+1} = \beta_p \cdot e^{k\cdot(T_s - R_s)}$$

HAE (Head Activation Enhancement)

- L_{Reg} just focuses on increasing the sparsity (usually makes all output prone to 0). we design a HAE Loss to activate more heads to avoid some experts always being asleep

$$\mathcal{L}_{HAE} = e^{2\cdot(R_s - T_s)} - 1 \quad \text{if } R_s > T_s.$$

- L_{HAE} helps some heads to be more active when they fall into a state of zero.

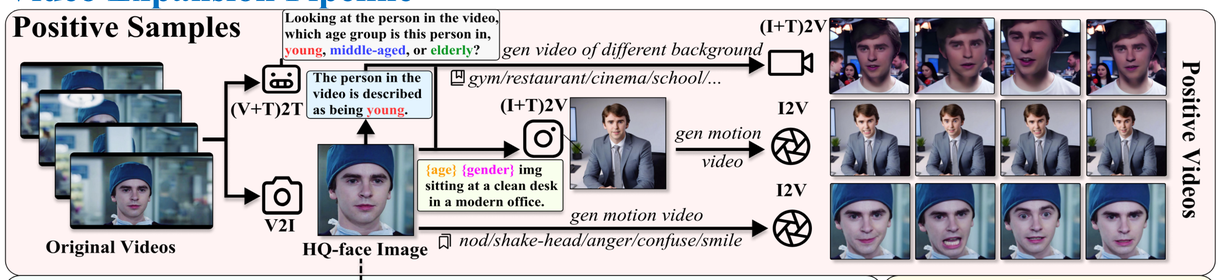
Final Loss

$$\mathcal{L} = \mathcal{L}_{LM} + \mathcal{L}_{SPR} + \mathcal{L}_{HAE}.$$

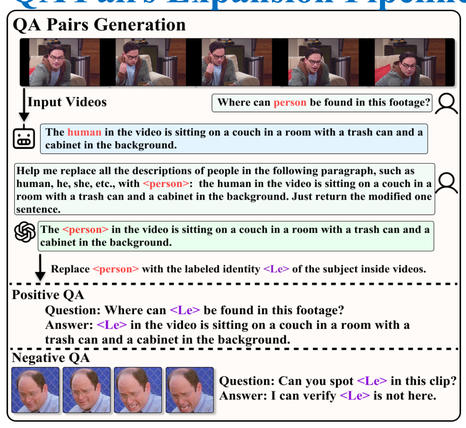
L_{LM} is the cross-entropy loss.

Data Pipeline and Experimental Results

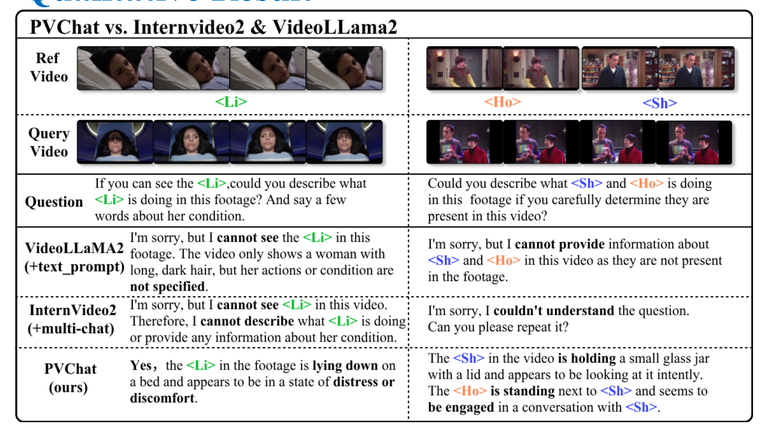
Video Expansion Pipeline



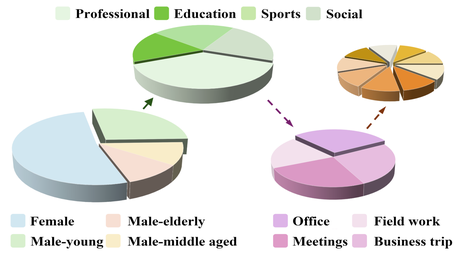
QA Pairs Expansion Pipeline



Qualitative Result

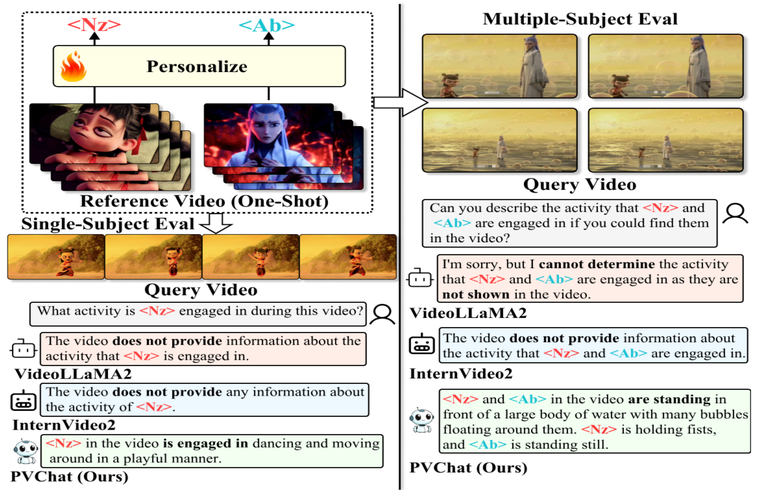


Prompt Library

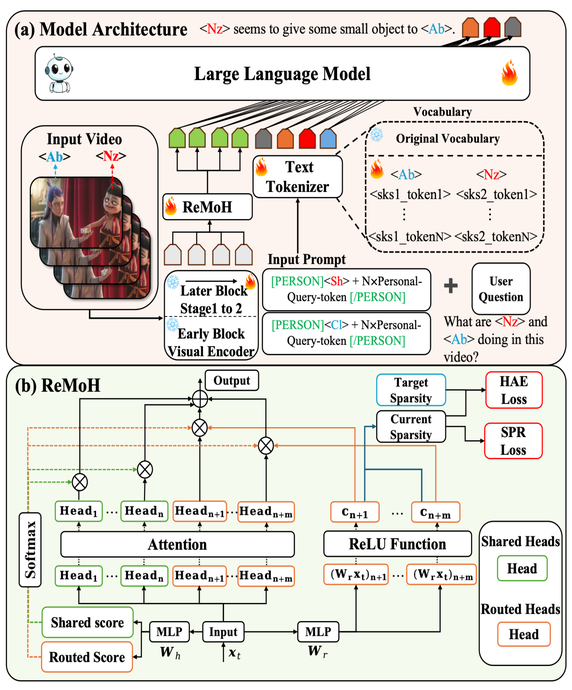


Quantitative Result

Model Type	Acc↑	BLEU↑	BS↑	ES↑	DC↑
InternVideo2 [50]	0.342	0.046	0.875	3.041	1.812
VideoLLaMA2 [5]	0.470	0.082	0.890	3.012	3.301
PVChat (**Ours**)	**0.901**	**0.562**	**0.952**	**4.940**	**4.201**



Model and ReMoH (ReLU Routing Mixture-of-Heads Attention)



- **Motivation:**

Top-k Mixture-of-Heads (MoH) routes tokens to a fixed number of heads with a non-differentiable choice, which (i) limits learning of subject-specific signals (ii) forces a brittle fixed-k (iii) can waste compute on irrelevant heads. Personalized video QA needs adaptive, fully differentiable head selection.

- **Multi-Head attention**

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V},$$

$$\text{where } \boldsymbol{Q} = \boldsymbol{X}_1\boldsymbol{W}_Q, \boldsymbol{K} = \boldsymbol{X}_2\boldsymbol{W}_K, \boldsymbol{V} = \boldsymbol{X}_2\boldsymbol{W}_V$$

$$\text{MHA}(\boldsymbol{X}_1,\boldsymbol{X}_2) = \text{Concat}\left(\boldsymbol{H}^1,\boldsymbol{H}^2,\ldots,\boldsymbol{H}^h\right)\boldsymbol{W}_O,$$

$$\boldsymbol{H}^i = \text{Attention}\left(\boldsymbol{X}_1\boldsymbol{W}_Q^i, \boldsymbol{X}_2\boldsymbol{W}_K^i, \boldsymbol{X}_2\boldsymbol{W}_V^i\right),$$

$$\text{MHA}(\boldsymbol{X}_1,\boldsymbol{X}_2) = \sum_{i=1}^{h} \boldsymbol{H}^i\boldsymbol{W}_O^i.$$

- **ReMoH**

$$\text{ReMoH}(\boldsymbol{X}_1,\boldsymbol{X}_2) = \textstyle\sum_{i=1}^{h} s_i \boldsymbol{H}^i\boldsymbol{W}_O^i,$$

$$s_i = \begin{cases} \alpha_1, & \text{if } 1 \le i \le n, \\ \alpha_2\,\text{ReLU}(\boldsymbol{W}_r\boldsymbol{x}_t)_i, & \text{if } n < i \le n+m \end{cases}$$

$$[\alpha_1,\alpha_2] = \text{Softmax}(\boldsymbol{W}_h\boldsymbol{x}_t)$$