PVChat: Personalized Video Chat with One-Shot Learning

Yufei Shi[1,5†], Weilong Yan[2†], Gang Xu[4], Yumeng Li[3], Yucheng Chen[1,5],
Zhenxi Li[1,5], Fei Yu[4], Ming Li[4✉], Si Yong Yeo[1,5✉]

[1]MedVisAI Lab [2]National University of Singapore [3]Nankai University
[4]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[5]Lee Kong Chian School of Medicine, Nanyang Technological University

Abstract

Video large language models (ViLLMs) excel in general video understanding, e.g., recognizing activities like talking and eating, but struggle with identity-aware comprehension, such as "Wilson is receiving chemotherapy" or "Tom is discussing with Sarah", limiting their applicability in smart healthcare and smart home environments. To address this limitation, we propose a one-shot learning framework PVChat*, the first personalized ViLLM that enables subject-aware question answering (QA) from* a single video for each subject*. Our approach optimizes a Mixture-of-Heads (MoH) enhanced ViLLM on a synthetically augmented video-QA dataset, leveraging a progressive image-to-video learning strategy. Specifically, we introduce an automated augmentation pipeline that synthesizes identity-preserving positive samples and retrieves hard negatives from existing video corpora, generating a diverse training dataset with four QA types: existence, appearance, action, and location inquiries. To enhance subject-specific learning, we propose a ReLU Routing MoH attention mechanism, alongside two novel objectives: (1) Smooth Proximity Regularization for progressive learning through exponential distance scaling and (2) Head Activation Enhancement for balanced attention routing. Finally, we adopt a two-stage training strategy, transitioning from image pre-training to video fine-tuning, enabling a gradual learning process from static attributes to dynamic representations. We evaluate* PVChat *on diverse datasets covering medical scenarios, TV series, anime, and real-world footage, demonstrating its superiority in personalized feature understanding after learning from a single video, compared to state-of-the-art ViLLMs. Code is PVChat.*

1. Introduction

Recent advancements in Video Large Language Models (ViLLMs) have showcased impressive capabilities across a wide range of tasks, *e.g.,* video captioning, temporal action localization, and video question answering [5, 32, 50, 51]. These models, trained on vast datasets, exhibit broad knowledge across multiple domains, such as medical assistance [17] and autonomous driving [56], significantly extending the functional scope of ViLLMs.

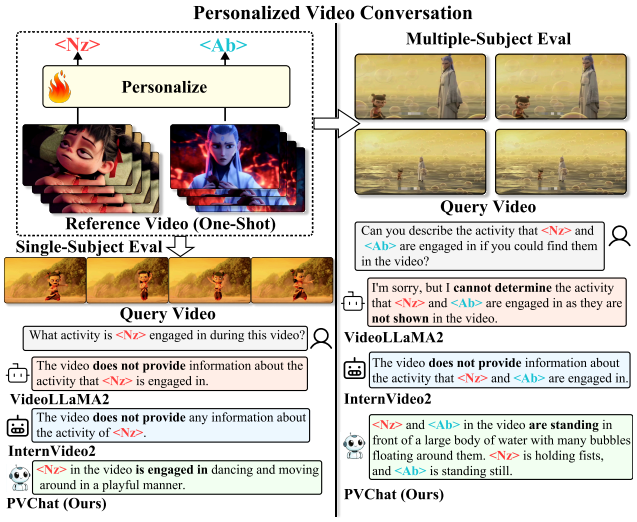

Figure 1. Examples of PVChat's ability with one-shot learning (e.g., <Nz>and <Ab>). PVChat can answer questions about the personalized information correctly while other models [5, 51] fail.

Despite these achievements, current ViLLMs are largely confined to general-purpose content understanding and struggle in personalized scenarios that require distinguishing specific individuals. For instance, when analyzing a video featuring a person named "Alex", even state-of-the-art models fail to consistently recognize "Alex" in novel contexts, such as identifying him from others or describing his specific actions. This limitation arises because existing training data predominantly focus on general knowledge rather than learning identity-specific information about an individual. As a result, ViLLMs are ill-equipped for real-world applications that demand personalized comprehension, *e.g.,* smart

† Equal contribution. ✉ Corresponding authors.

home [61], intelligent healthcare [15] and human-computer interaction [6].

To enhance the ability of LLMs in capturing personalized cues, recent research has focused on personalized image understanding [35, 38, 39]. For example, Yo-LLaVA [35] acquires the capability to encode a customized subject into a collection of latent tokens by utilizing a small number of sample images representing the subject. Other studies [38, 39] construct specialized training datasets curated with expert annotations to enable personalized image recognition. While these methods have demonstrated promising results, they are inherently limited to static image-based understanding and cannot model dynamic, temporally evolving visual cues present in videos. Given that videos contain significantly richer personalized information, including motion patterns, interaction dynamics, and contextual dependencies, there is a pressing need for solutions that enable identity-aware comprehension in video-based contexts.

To address these limitations, we propose *PVChat*, the first personalized video understanding LLM capable of learning individual characteristics and supporting subject-specific question answering from a single reference video. To enable one-shot learning, we propose a data augmentation pipeline that synthesizes high-quality personalized samples. It generates identity-preserving positives while incorporating visually similar negatives for enhanced discrimination. The process begins with facial attribute extraction and demographic categorization, followed by high-fidelity video synthesis using ConsisID [57] and PhotoMaker [21]. Hard negatives are introduced by retrieving similar faces from Laion-Face-5B [60] and CelebV-HQ [62] with corresponding video synthesis. Finally, question-answer pairs covering four key categories—subject existence, appearance, action, and location—are generated via InternVideo2 [51] and ChatGPT-4o to ensure linguistic coherence and personalization.

To improve subject-specific feature learning, we introduce a ReLU Routing Mixture-of-Heads (ReMoH) attention mechanism, which replaces conventional softmax-based and top-k head selection with a ReLU-driven dynamic routing strategy, enabling smooth and scalable training. Additionally, we propose two novel optimization objectives: Smooth Proximity Regularization, which promotes progressive learning via exponential distance scaling, and Head Activation Enhancement, which balances shared and routed attention heads, effectively mitigating gradient explosion and inactive heads in multi-head attention mechanisms. To optimize our ReMoH-inspired *PVChat* under limited data conditions, we adopt a two-stage training strategy transitioning from static image learning to video temporal modeling. In the first stage, the model is trained on images with a simple summary task to capture static identity attributes. In the second stage, it is fine-tuned on video QA tasks to develop dynamic action recognition and multi-person interaction capabilities. This progressive learning framework enables a structured transition from static representations to complex spatiotemporal reasoning. We verify *PVChat* across a diverse range of personalized scenarios, including healthcare, TV series, anime, and real-world scenes, requiring the recognition of one, two, or three individuals. Experimental results demonstrate that *PVChat* achieves state-of-the-art performance in individual information comprehension and identity-aware reasoning from a single reference video.

In summary, our main contributions of this work are listed as follows:

- We introduce *PVChat*, the first ViLLM that extends LLMs with personalization capability from a single reference video. This enables personalized video understanding and user-specific question-answering with only one-shot learning, laying a strong foundation for future individual video comprehension.
- We develop a systematic video augmentation and QA generation pipeline from a single reference video. The key is to exploit off-the-shelf toolboxes to generate identity-preserving positives and synthesize hard negatives for confusingly similar faces. To support the research, we construct a diverse dataset containing 6 individual scenarios, 304 original videos, 2,304 extended generated videos with over 30,000 QA pairs, which will be publicly released to facilitate future research.
- We design the ReLU Routing Mixture-of-Heads module to enhance the effective extraction of individual-specific characteristics from videos. Additionally, we introduce Smooth Proximity Regularization and Head Activation Enhancement to improve training stability and attention head activation.

2. Related Works

2.1. Image Large Language Models

The remarkable language comprehension and reasoning capabilities of Large Language Models (LLMs) [1, 3, 46, 48] have spurred significant interest in extending them into multimodal understanding. Pioneering works demonstrate different approaches for vision-language alignment: Flamingo [48] processes interleaved image-text inputs through cross-attention layers for diverse multimodal tasks, while BLIP-2 [16] employs a Querying Transformer (Q-Former) to bridge frozen visual encoders with LLMs. Simpler architecture like LLaVA [24] achieves effective feature alignment using lightweight MLP projectors. Recent research systematically explores critical components of Multimodal LLMs (MLLMs), including dynamic high-resolution processing strategies [4, 26], instruction-tuning methodologies [18, 29], and optimal visual encoder selection [47, 53]. Comprehensive design space analyses from MM1 [33] and Idefics2 [14] further establish foundational principles for

MLLM development.

2.2. Video Large Language Models

With progress in image-based multimodal language models (MLLMs), video understanding research has garnered great attention. Video-ChatGPT [31] and Vally [30] aggregate temporal features into compact tokens, while Video-LLaVA [22] unifies image-video representations through aligned projections before LLM integration. Video-Teller [23] underscores modality alignment in pre-training objectives, whereas PLLaVA [55] examines feature pooling impacts on video Question-Answering tasks. For extended sequences, LLaMA-VID [19] encodes frames with dual tokens to balance efficiency and information retention, while MovieChat [45] employs memory optimization for processing ultra-long videos (>10K frames). Weng *et al.* [54] enhance local segment features by injecting global semantics through hierarchical fusion mechanisms. These approaches collectively address spatiotemporal modeling challenges across varied input scales. However, these models are still incapable of understanding videos with personalization.

2.3. Personalized Large Language Models

Personalization in generative models exhibits distinct implementations across domains. For image generation, existing approaches focus on pixel-level subject fidelity through either concept token optimization [8, 13] or model parameter adaptation [41]. In natural language processing contexts, personalization typically involves shaping LLMs' behavioral patterns (*e.g.,* , conversational style) via prompt engineering or metadata-driven retrieval mechanisms [28, 36, 59]. Recent personalized modeling approaches employ parameter-efficient adaptation through few-shot learning. MyVLM [2] implements a Q-former-based architecture with trainable projection heads for concept extraction and visual feature augmentation, whereas Yo'LLaVA [34] extends the LLaVA framework by integrating novel concepts as specialized tokens within the LLM's embedding space. To the best of our knowledge, our model is the first to support videos input while others only accept images as inputs.

3. *PVChat*

3.1. Overview

To achieve video personalization understanding, we first establish a promising and systematic data augmentation pipeline in 3.2 to deal with the severe lack of diverse video data for individuals. To enhance the subject-specific learning, we propose a ReLU Routing MoH Attention mechanism in 3.3, alongside two novel objectives: (1) Smooth Proximity Regularization (SPR) for progressive learning through exponential distance scaling; (2) Head Activation Enhancement (HAE) for balanced attention routing. Finally, a two-stage training strategy is adopted in 3.4 to progressively learn from static features to dynamic features of the target.

3.2. Data Collection

To address the challenge of limited training data and achieve one-shot learning, we propose a novel and comprehensive data augmentation pipeline as illustrated in Fig.2, which starts from two key principles: 1) enabling automated personalized data collection directly from a wide range of natural videos; 2) ensuring that the generated personalized data encompasses different scenes, actions, and contextual elements while preserving the identity information.

Identity-Preserving Positives Generation. Maintaining the identity information is key to obtaining diverse data of a specific individual. To achieve that, our pipeline starts with facial extraction from original videos using DeepFaceLab [37]. When multiple faces are present, we carefully design a two-stage approach: first, we extract facial embeddings via FaceNet [42], then apply DBSCAN [7] clustering to differentiate between subjects. After that, the faces are evaluated using several metrics, including Eye Aspect Ratio (EAR) to determine eye openness (with a threshold of 0.25), facial orientation, image clarity, and presence of facial landmarks. These metrics allow us to identify the highest-quality facial image (HQ-face) for subsequent steps.

To further enhance the stability of identity preservation, the characteristics of individuals in the videos are determined using InternVideo2 [51], which are categorized by gender (male/female) and age group (young/middle-aged/elderly) based on the video content and our designed prompt. With these demographic characteristics, we can more accurately preserve the identity, and employ ConsisID [57] with our extensive prompt library Fig. 5 to generate a wide range of videos in different scenarios (e.g., gym, restaurant, school, etc) with identity consistency. Although videos from ConsisID [57] contain rich content and we extract multiple types of identity information, their ID consistency is still poor, leading to a decline in data quality.

With the aim of incorporating more ID consistent data, we utilize PhotoMaker [21] with the identified characteristics to generate additional high-fidelity images of the subject across different environments, and adopt LivePortrait [9] to animate these images to generate high-identity-consistency videos with a facial motion and expression library (including nodding, head-shaking, questioning, anger, and smiling). Meanwhile, we also directly input the HQ-face image to LivePortrait [9], generating the motion video. These videos exhibit strong ID consistency but include simple contents, complementing the data from ConsisID [57].

Hard Negatives Retrieval. Avoiding the misidentification of personalized subjects is crucial for personalized video chatting. As observed in previous VLM research [35] and our video chat experiments, training with only positive data

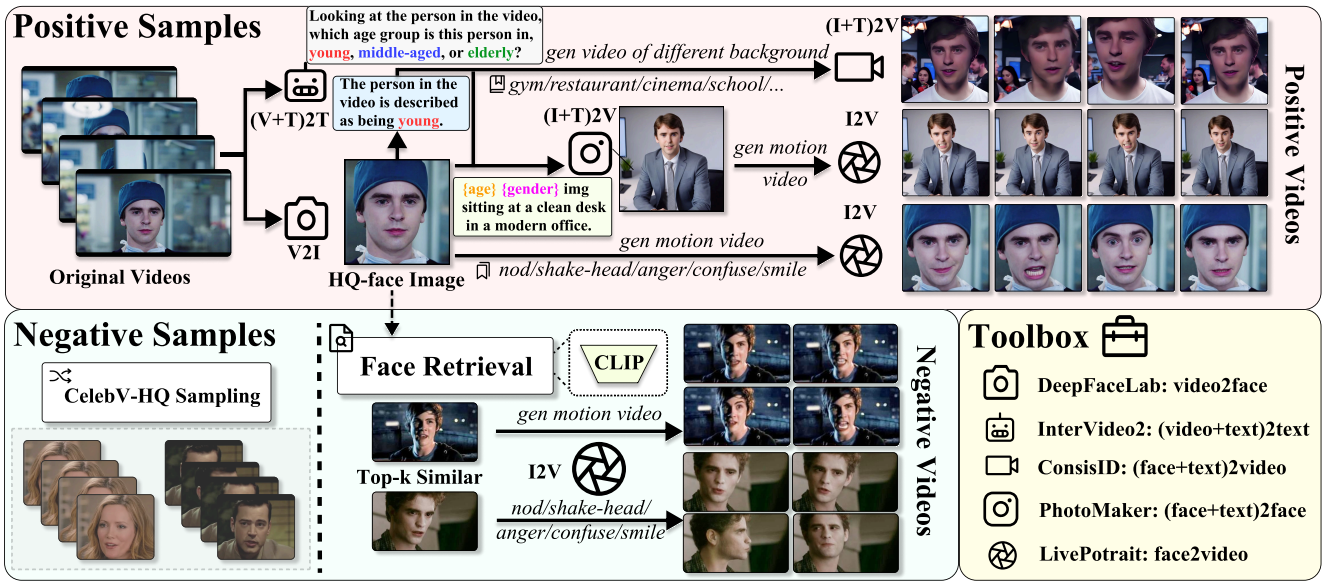

Figure 2. The systematic data collection pipeline. For positive data collection, the original videos are processed by DeepFaceLab [37] for high-quality face and InterVideo2 [51] for demographic characteristics, which boost identity preservation. ConsisID [57] and LivePortrait [9] with PhotoMaker [21] utilize the identity information to generate videos of various background or different motion/expression, respectively. For model's robust perception, hard negative samples are selected from either similar face retrieval to generate negative videos, or sampled from the CelebV-HQ dataset [62]. These negative samples guarantee the model's accurate recognition of both identity and content.

can lead to the model losing its ability to recognize personalized subjects. Considering these factors, we introduce challenging negative samples by retrieving top-k visually similar faces from the Laion-face-5b dataset [60] utilizing CLIP [40]. Similarly, these facial images are animated by LivePortrait [9] to serve as negative data, which enables the model to learn to robustly identify the personalized subjects. In addition, these negative samples also contain relatively simple movements and other content, which inspires us to supplement them with 30 randomly selected videos from the CelebV-HQ dataset [62] that possesses rich and high-resolution content. Combining them together, the personalized video chat model is equipped with robustness to both subjects and content.

Question-Answer Pairs Generation. Having various positive and negative samples, the final stage involves generating question-answer pairs using InternVideo2 [51] across four categories: existence verification, appearance description, location identification, and action recognition. To enhance the linguistic naturalness, we refine these responses using ChatGPT-4o, converting generic human references into subject-specific references. The QA generation is illustrated in Fig. 3.

Through this comprehensive pipeline, each input natural video is transformed into 81 different videos accompanied by 1,455 question-answer pairs, significantly expanding our training dataset while maintaining subject identity consistency and guaranteeing the model's robustness.

3.3. ReLU Routing Mixture-of-Heads Attention

Multi-Head attention. We begin by reviewing the standard multi-head attention mechanism [49] which is based on scaled dot-product attention. For T_1 input tokens $\boldsymbol{X}_1 \in \mathbb{R}^{T_1 \times d_{in}}$ where d_{in} is the input dimension, and T_2 tokens $\boldsymbol{X}_2 \in \mathbb{R}^{T_2 \times d_{in}}$, attention is conducted as follows:

$$\begin{aligned} &\text{Attention}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V}, \\ &\text{where } \boldsymbol{Q} = \boldsymbol{X}_1\boldsymbol{W}_Q, \boldsymbol{K} = \boldsymbol{X}_2\boldsymbol{W}_K, \boldsymbol{V} = \boldsymbol{X}_2\boldsymbol{W}_V. \end{aligned} \tag{1}$$

To enhance the representation of the attention layer, Vaswani [49] proposed to use multiple attention heads to operate on the projection of the input. The transformer computes h different projection of $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$, and the concatenation form of multi-head attention is

$$\begin{aligned} \text{MHA}\left(\boldsymbol{X}_1, \boldsymbol{X}_2\right) = &\text{ Concat}\left(\boldsymbol{H}^1, \boldsymbol{H}^2, \ldots, \boldsymbol{H}^h\right)\boldsymbol{W}_O, \\ \boldsymbol{H}^i = &\text{ Attention}\left(\boldsymbol{X}_1\boldsymbol{W}_Q^i, \boldsymbol{X}_2\boldsymbol{W}_K^i, \boldsymbol{X}_2\boldsymbol{W}_V^i\right), \end{aligned} \tag{2}$$

where $\boldsymbol{W}_Q^i, \boldsymbol{W}_K^i, \boldsymbol{W}_V^i$ represent the i_{th} projection matrices of the query, key and value. $\boldsymbol{W}_O$ is the final projection matrix, which is decomposed by $\boldsymbol{W}_O = \left[\boldsymbol{W}_O^1, \boldsymbol{W}_O^2, \ldots, \boldsymbol{W}_O^h\right]$. The summation form is as follows:

$$\text{MHA}\left(\boldsymbol{X}_1, \boldsymbol{X}_2\right) = \sum_{i=1}^{h} \boldsymbol{H}^i \boldsymbol{W}_O^i. \tag{3}$$

3.3.1. Heads as Experts with ReLU Routing

Motivation of ReMoH. Recent study Mixture-of-Heads [12] targets to allow each token to select the Top-k relevant

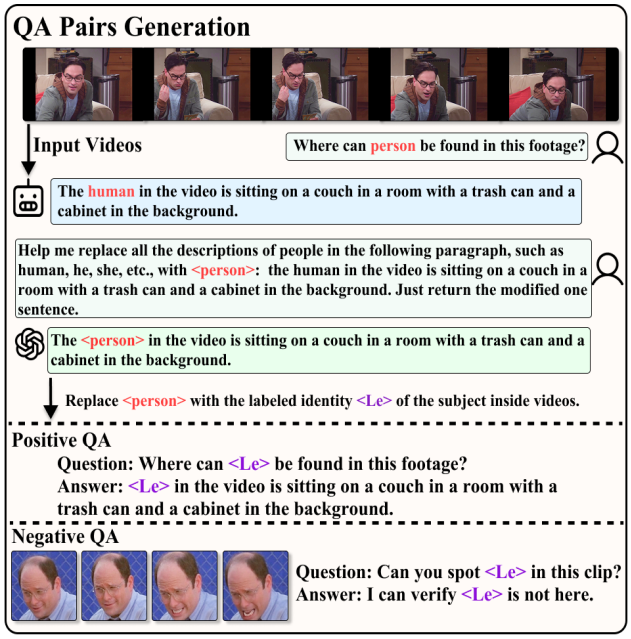

Figure 3. We illustrate the process of automatically generating question-answer pairs using InternVideo2 [51] and ChatGPT [1]. A positive and a negative sample are shown at the bottom.

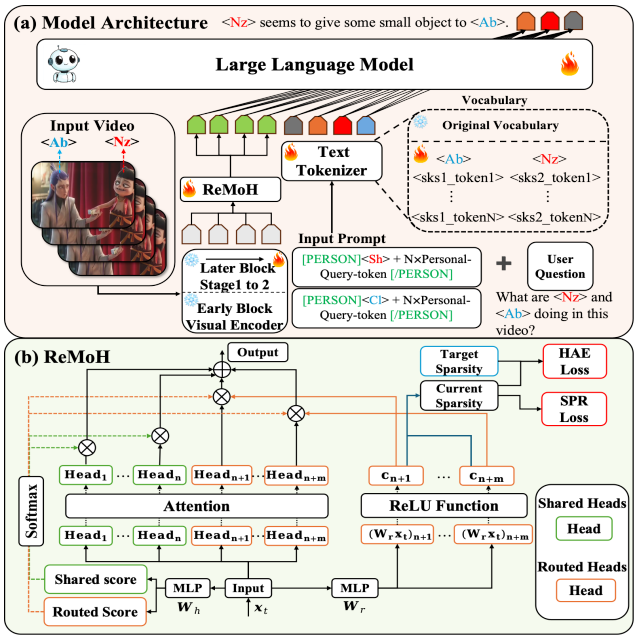

Figure 4. (a) The training pipeline of our method. (b) The proposed ReMoH technique for better specialized characteristics learning.

heads, improving inference efficiency without sacrificing accuracy or increasing the parameters, compared with traditional multi-head attention. Moreover, MoH [12] finds it can learn domain-specific information, which is consistent with our personalized video understanding with only one-shot learning. However, the vanilla Top-k choice of MoH limits the performance and domain-specific information learning, which is because: 1) Top-k choice is not fully differentiable for training; 2) The choice of heads is not adaptive and flexible, and interferes with the domain-specific knowledge learned by expert heads. Thus, inspired by [52], our ReLU Routing Mixture-of-Heads Attention (ReMoH) aims at enhancing MoH with much more efficient and adaptive learning especially for the personalized subject's information, with only an increase of 2 MLP parameters. The detailed design for ReMoH can be seen in Fig. 4 (b).

ReMoH with Head and ReLU Routers. Specifically, ReMoH consists of h attention heads $\boldsymbol{H} = \{H^1, H^2, \ldots, H^h\}$ and uses a head routers with a ReLU router to modulate the output of different expert heads. Formally, given input tokens $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, the output of ReMoH is the weighted sum of outputs of the heads:

$$\mathrm{ReMoH}\left(\boldsymbol{X}_1, \boldsymbol{X}_2\right) = \textstyle\sum_{i=1}^{h} s_i \boldsymbol{H}^i \boldsymbol{W}_O^i, \tag{4}$$

where s_i is the score corresponding to H^i. In the attention mechanism, some attention heads capture common knowledge across different contexts, such as grammatical rules in language, while others focus on context-specific knowledge [43]. Following MoH [12], we divide the heads into shared heads and routed heads, where shared heads are always activated, and routed heads will only be activated if the ReLU routing score is non-zero. Specifically, the score for each head is defined as:

$$s_i = \begin{cases} \alpha_1, & \text{if } 1 \le i \le n, \\ \alpha_2 \,\mathrm{ReLU}\left(\boldsymbol{W}_r \boldsymbol{x}_t\right)_i, & \text{if } n < i \le n+m, \end{cases} \tag{5}$$

where n and m are the number of shared heads and routed heads, respectively. $\boldsymbol{W}_r \in \mathbb{R}^{m \times d_{in}}$ denotes the projection matrix of the ReLU router. α_1 and α_2, which come from the shared router, will be used to balance the contributions of the shared and routed heads, and are defined as:

$$[\alpha_1, \alpha_2] = \mathrm{Softmax}\left(\boldsymbol{W}_h \boldsymbol{x}_t\right). \tag{6}$$

With the utility of ReLU router to choose specific routed heads to be activated, the decision process becomes fully differentiable and offers greater flexibility. This boosts the learning of the personalized video information, avoiding a fixed number of activated experts to limit the performance or cause computational redundancy. A more detailed analysis can be found in our ablation study.

3.3.2. Controlling Sparsity Strategy

One important issue found in [52] is that most experts are always activated during training. Our ReMoH aims to control computation cost by managing the sparsity of the ReLU output, targeting a sparsity of $T_s = (1 - \frac{b}{m})$, where b is the number of routed heads we set for encouraging the activation. To encourage the sparsity of activation, which helps to

reduce computational cost and boost specific-domain learning, we propose the Smooth Proximity Regularization (SPR) Loss to make the training more flexible:

$$\mathcal{L}_{SPR} = \beta_p \cdot \mathcal{L}_{Reg}, \tag{7}$$

where β_p is a step-p adaptive weight which adjusts the strength of L_{Reg}. Following prior work [20, 44], to effectively encourage sparsity by penalizing the average activation values, L_{Reg} always uses the L_1 regularization:

$$\mathcal{L}_{Reg} = \left\| \frac{1}{n}(\boldsymbol{W_r x_t}) \right\|, \tag{8}$$

$$\beta_{p+1} = \beta_p \cdot e^{k\cdot(T_s - R_s)}, \tag{9}$$

where R_s is the current sparsity and $R_s = 1 - \frac{1}{n}(\boldsymbol{W_r x_t})$, k is a scaling factor to control the change of the step-wise weight. Such a design will make the training process more flexible, more stable, and with less extreme loss.

However, L_{Reg} just focuses on increasing the sparsity, forcing the model to be more sparsely activated, which usually makes all output prone to 0. Considering this, we design a Head Activation Enhancement (HAE) Loss to activate more heads to avoid some experts always being asleep:

$$\mathcal{L}_{HAE} = e^{2\cdot(R_s - T_s)} - 1 \quad \text{if } R_s > T_s. \tag{10}$$

$\mathcal{L}_{HAE}$ helps some heads to be more active when they fall into a state of zero. Combining the SPR and HAE, a balanced activation ratio can be guaranteed. Our final training objective is designed with the language-model used cross-entropy loss $\mathcal{L}_{LM}$:

$$\mathcal{L} = \mathcal{L}_{LM} + \mathcal{L}_{SPR} + \mathcal{L}_{HAE}. \tag{11}$$

3.4. Training Pipeline

We implement a two-stage training strategy, transitioning from image-summary pretraining to video-QA fine-tuning, teaching the model to progressively learn both static and dynamic subject-specific features. This strategy adapts our *PVChat* model from static image understanding to dynamic video reasoning. The architecture is shown in Fig. 4(a).

Stage 1: Image Understanding. In the first stage, the model focuses on learning static subject-specific knowledge from separate images, where we extract the first frame of each video as input. To preserve the pre-trained visual encoder's capability, we freeze the visual encoder and only train the ReMoH components, by employing Low-Rank Adaptation (LoRA) [10] to efficiently fine-tune the Mistral-7B-Instruct-v0.3 language model [11], which significantly reduces trainable parameters while maintaining performance.

The QA pairs used in this stage primarily consist of existence verification (e.g., "Is <sks>visible in this video?") and attribute description questions (e.g.,"What is <sks>wearing in this video?"). These questions guide the model to focus on static subject-specific features, which helps the model to do personalized image chat.

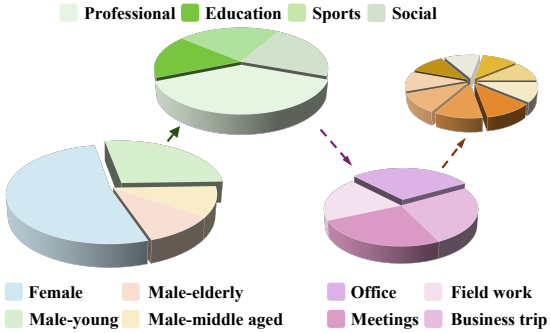

Figure 5. The hierarchical structure of our prompt library, which is carefully divided into four levels, such as gender, age, and scenarios, and provides different descriptions according to the specific subject.

Stage 2: Video Reasoning. In the second stage, we extend the model to capture temporal dynamics and environmental interactions. We selectively train several later blocks in the visual encoder to enhance cross-frame feature integration in the personalized scenarios. This modification enables the model to recognize how subjects interact with their environments and what the subject is doing over time.

Here, we expand our QA dataset with action recognition questions (e.g., "What movements or actions does <sks>perform here?") and location identification questions (e.g., "Can you describe <sks>'s location relative to others?"). Utilizing both positive and negative samples for training, our *PVChat* model becomes robust and precise.

4. Experiments

4.1. Settings

Implementation Details. Our LLM backbone is based on the Mistral-7B-Instruct-v0.3 [11]. All models are trained on one NVIDIA L20 GPU. The first stage of training takes one epoch, and the second stage takes seven epochs. We set the batch size to 2, the learning rates of token embedding, ReMoH, LLM Lora to 1×10^{-4}, 1×10^{-5}, and 1×10^{-5}, training time needs $3h$, video resolution is 1080×1920.For visual encoder each video will evenly sample 8 frames.

Datasets. To establish the dataset, we collect six different scenarios, including (Friends(6), Good Doctor(5), Ne Zha(2), doctor(3), patient(3), Big Bang(6)) 25 characters (including doctor, patient, cartoon character, engineer, professor, etc.), and collect 304 original videos and 2,304 extended videos with over 30,000 QA pairs. 300 prompt descriptions for different scenarios are designed for different ages and genders. The structure and distribution of our dataset can be found in Fig. 5. For each character, we randomly select one video for the training set as our refer video and the other for the evaluation set as query video. We show more details and

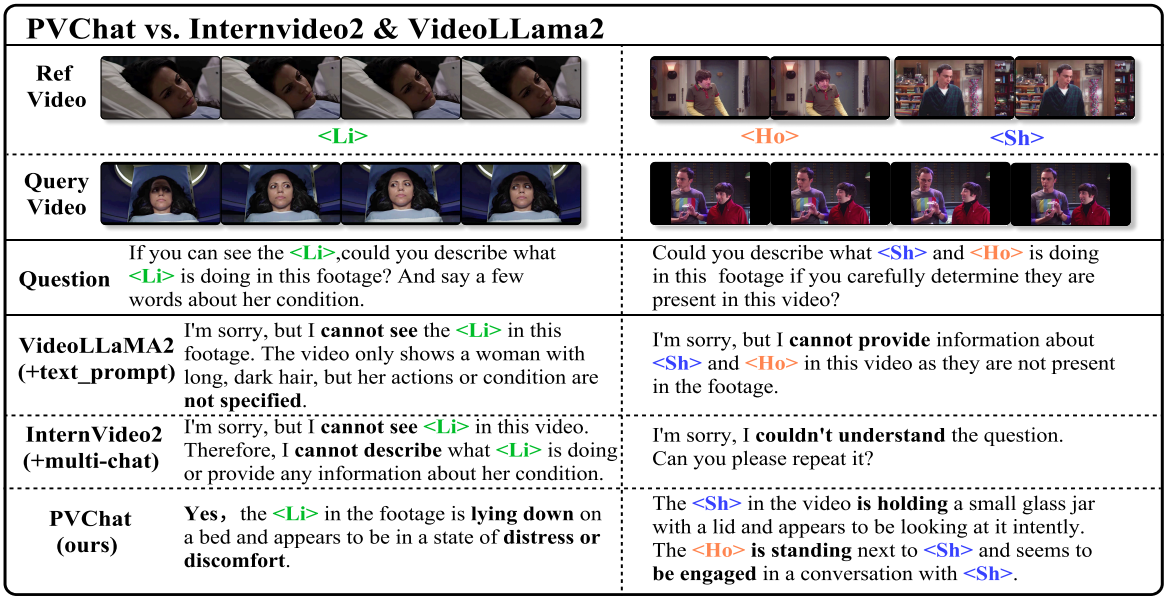

Figure 6. Examples of PVChat's ability with a learned video (e.g., a man named <Sh>and another man named <Ho>). PVChat can recognize and answer questions about the personalized concept in various scenarios, such as medical scenarios (left) and TV series (right).

examples in our supplementary.

Evaluation Metrics across Multiple Datasets. Following the evaluation of LLaVA [25] and Yo'LLaVA [35], we rigorously assess *PVChat* with state-of-the-art ViLLMs using five complementary metrics across four question categories: existence, appearance, action, and location queries. The final results presented represent the averaged performance across all datasets, offering a holistic view of model capabilities.

Our evaluation metrics incorporate: (1) Accuracy (Acc) to measure the correctness of the presence or absence of subjects; (2) BLEU and BERTScore (BS) to quantify the textual similarity between model's response and the ground truth; (3) Entity Specificity (ES) to assess if responses contain personalized information; (4) Descriptive Completeness (DC) to measure the quality of details inside responses. More details about the evaluation metrics can be found in the supplementary. For appearance, action, and location queries, we employ BLEU, BS, ES, and DC to comprehensively evaluate both semantic accuracy and conversational quality of generated responses.

4.2. Comparison with State-of-the-Art Models

Quantitative Results. To verify PVChat's effectiveness, we present the quantitative results in Tab. 1, which is compared with state-of-the-art ViLLMs, including InternVideo2 [51], VideoLLaMA2 [5].

The result demonstrates that our PVChat exhibits a strong ability for personalized video understanding, offering high-accuracy and comprehensive responses. Especially for the negative sample, another model always replies that the object character is existent and no real understanding of the characteristic of the object. Since these models do not support multi-video input in a single-round conversation, we are not able to input the reference video and query video simultaneously. To ensure fairness, we find that InternVideo2 [51] supports the multi-round conversation, allowing us to use the reference video in the first round conversation, and input our query video in the second round. VideoLLaMA2 [5] does not support the multi-round conversation, so we input a detailed description of the reference video, and then test the query video using the character's detailed description.

Model Type	Acc↑	BLEU↑	BS↑	ES↑	DC↑
InternVideo2 [51]	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**

Table 1. Quantitative evaluation of our method against state-of-the-art methods [5, 51]. Compared with these SOTA models, our model exhibits superior performance across five metrics.

Qualitative Analysis. We conduct qualitative assessments on some challenging scenarios, including single-subject and multi-subject settings in various interactive environments and actions in Fig. 6. The result shows that our *PVChat* outperforms other models [5, 51] in different settings, demonstrating our superior personalized video understanding ability. Due to the shortage of personalized training and subject-specific data, previous models cannot concisely capture personalized information and show reasonable responses with correct identity.

4.3. Ablation Study

Analysis of ReMoH. To explore the impact of ReMoH, we conduct experiments as shown in Tab. 2 and visualize the

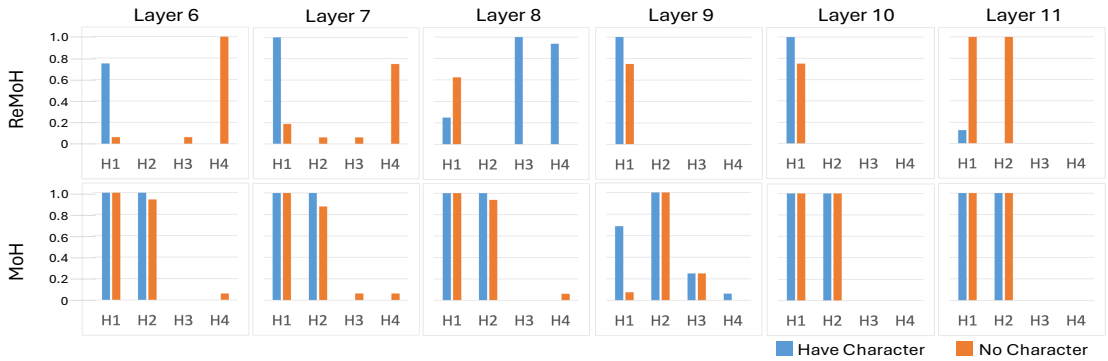

Figure 7. The comparison of expert heads activation between MoH [12] and ReMoH in different layers, where H_i represents the i^{th} head. Orange refers to the video without the target individual, while blue represents the video having the character.

Method	Acc↑	BLEU↑	BS↑	ES↑	DC↑
Baseline	0.733	0.550	0.904	4.735	4.142
Baseline + MoH [12]	0.813	0.558	0.926	4.939	4.191
Baseline + ReMoH	**0.901**	**0.562**	**0.952**	**4.940**	**4.201**

Table 2. Ablations on the proposed ReMoH, where ReMoH significantly outperforms MoH in Acc.

heads of expert activation ratio in Fig. 7. We choose the typical transformer architecture Q-former [58] with the LLM backbone [11] as our baseline. Compared with the baseline and incorporating the typical MoH [12], our ReMoH results in respective improvements of 18.6%, 9.8% on the Acc, where our ReMoH significantly enhances the model's capability of personalizing understanding and chatting.

In Fig. 7, we find that ReMoH effectively allocates some certain heads of expert to extract the video feature about the target character, while MoH [12] almost evenly distributes expert heads to learn the knowledge. For example, if using MoH [12], in most layers, no matter the character appears in the video or not, the first and second heads are always activated for video reasoning, verifying that these heads of expert actually do not learn about specific information. On the contrary, with the utility of ReMoH, in most layers it is observed that the activation of expert heads undergoes significant changes from no characters to having characters. This pattern supports that the model can learn personalized knowledge with ReMoH better, improving the QA performance related to specific identity, as presented in Tab. 2.

Influence of SPR and HAE. We conducted ablations to verify the effectiveness of SPR and HAE losses. As shown in Tab. 3, the introduction of SPR successfully smooths loss fluctuations and enhances training stability compared with MoH [12]. However, models with only SPR suffer from very low expert head activation ratio, limiting personalized feature extraction. Adding HAE loss alongside SPR, the head activation diversity increases significantly, allowing the model to capture more nuanced personalized representations.

Contribution of Each Type of Data. We conduct an ablation on the dataset creation. Tab. 4 presents all metrics

Method	Activate Rate	loss↓	Acc↑	BLEU↑	BS↑	ES↑	DC↑
PVChat *w/o* SPR and MAE	–	nan	–	–	–	–	–
PVChat *w/o* MAE	0.217	0.085	0.746	0.555	0.926	4.913	4.112
PVChat	**0.552**	**0.028**	**0.901**	**0.562**	**0.952**	**4.940**	**4.201**

Table 3. Ablations on SPR and HAE losses. It verifies that SPR and HAE guarantee stability and enhance learning of the expert heads.

Data Type	Acc↑	BLEU↑	BS↑	ES↑	DC↑
one positive	0.464	0.417	0.905	4.826	3.947
+Negative	0.584	0.418	0.931	4.899	4.120
+ConsisID Positive	0.781	0.532	0.927	4.929	4.132
+LivePortrait Positive	**0.901**	**0.562**	**0.952**	**4.940**	**4.201**

Table 4. Ablations on the dataset collection, where combining all types of designed data, the model performs accurately and robustly.

for the personalization conversion. Using vanilla datasets (with only one positive sample) fails to perform well on personalized questions, always responding with "yes" to all questions. After injecting the negative samples, it could recognize the identity of the character to a certain degree, but still suffers from some complex questions about action and location. After training with the generated data from ConsisID [57] and LivePortrait [9], the model is capable of understanding the specific information more accurately and robustly.

Influence of Token Numbers. We set each question prompt corresponding to $N = 16$ tokens per character. As the token number increases, the metrics overall increase, but decrease when larger than 16. This is because too many tokens make the model hard to capture the characteristics of the subject. The detailed results can be found in our supplementary.

5. Conclusion

This paper introduces *PVChat*, the first personalized ViLLM that enables personalized subject learning from a single video. We collect comprehensive datasets over 2,300 videos, propose ReMoH for specific knowledge learning, and employ a two-stage training strategy from image pre-training to video fine-tuning, with SPR and HAE losses. Evaluations show the superiority of our model in personalized video QA, making it potential for smart healthcare and home scenarios.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 5

[2] Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries. In *European Conference on Computer Vision*, pages 73–91. Springer, 2024. 3

[3] Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024-09-18. 1,2. 2

[4] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 2

[5] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 7

[6] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 902–909, 2024. 2

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 3

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[9] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 3, 4, 8

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 6

[11] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 6, 8

[12] Peng Jin, Bo Zhu, Li Yuan, and Shuicheng Yan. Moh: Multi-head attention as mixture-of-head attention. *arXiv preprint arXiv:2410.11842*, 2024. 4, 5, 8

[13] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 3

[14] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37: 87874–87907, 2025. 2

[15] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 2

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[17] Jiajie Li, Garrett Skinner, Gene Yang, Brian R Quaranto, Steven D Schwaitzberg, Peter CW Kim, and Jinjun Xiong. Llava-surg: towards multimodal surgical assistant via structured surgical video learning. *arXiv preprint arXiv:2408.07981*, 2024. 1

[18] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2

[19] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 3

[20] Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, et al. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. *arXiv preprint arXiv:2210.06313*, 2022. 6

[21] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8640–8650, 2024. 2, 3, 4

[22] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 3

[23] Haogeng Liu, Qihang Fan, Tingkai Liu, Linjie Yang, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. Video-teller: Enhancing cross-modal generation with fusion and decoupling. *arXiv preprint arXiv:2310.04991*, 2023. 3

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 7

[26] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 2

[27] Jinhua Liu, Christian Desrosiers, Dexin Yu, and Yuanfeng Zhou. Semi-supervised medical image segmentation using cross-style consistency with shape-aware and local context constraints. *IEEE Transactions on Medical Imaging*, 43(4): 1449–1461, 2023.

[28] Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. You impress me: Dialogue generation via mutual persona perception. *arXiv preprint arXiv:2004.05388*, 2020. 3

[29] Yuan Liu, Zhongyin Zhao, Ziyuan Zhuang, Le Tian, Xiao Zhou, and Jie Zhou. Points: Improving your vision-language model with affordable strategies. *arXiv preprint arXiv:2409.04828*, 2024. 2

[30] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 3

[31] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3

[32] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1

[33] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer, 2024. 2

[34] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant. *Advances in Neural Information Processing Systems*, 37:40913–40951, 2025. 3

[35] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant. *Advances in Neural Information Processing Systems*, 37:40913–40951, 2025. 2, 3, 7

[36] Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*, 2022. 3

[37] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020. 3, 4

[38] Chau Pham, Hoang Phan, David Doermann, and Yunjie Tian. Personalized large vision-language models. *arXiv preprint arXiv:2412.17610*, 2024. 2

[39] Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. Personalized visual instruction tuning. *arXiv preprint arXiv:2410.07113*, 2024. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4

[41] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3

[43] Dai Shi. Transnext: Robust foveal visual perception for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17773–17783, 2024. 5

[44] Chenyang Song, Xu Han, Zhengyan Zhang, Shengding Hu, Xiyu Shi, Kuai Li, Chen Chen, Zhiyuan Liu, Guangli Li, Tao Yang, et al. Prosparse: Introducing and enhancing intrinsic activation sparsity within large language models, july 2024. *arXiv preprint arXiv:2402.13516*. 6

[45] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 3

[46] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[47] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2025. 2

[48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[50] Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*, 2023. 1

[51] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong

Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 1, 2, 3, 4, 5, 7

[52] Ziteng Wang, Jianfei Chen, and Jun Zhu. Remoe: Fully differentiable mixture-of-experts with relu routing. *arXiv preprint arXiv:2412.14711*, 2024. 5

[53] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer, 2024. 2

[54] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024. 3

[55] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 3

[56] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. 1

[57] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. *arXiv preprint arXiv:2411.17440*, 2024. 2, 3, 4, 8

[58] Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. Vision transformer with quadrangle attention. *arXiv preprint arXiv:2303.15105*, 2023. 8

[59] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018. 3

[60] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 2, 4

[61] Yaoyao Zhong, Mengshi Qi, Rui Wang, Yuhan Qiu, Yang Zhang, and Huadong Ma. Viotgpt: Learning to schedule vision tools towards intelligent video internet of things. *arXiv preprint arXiv:2312.00401*, 2023. 2

[62] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 2, 4