

PVChat: Personalized Video Chat with One-Shot Learning

-Supplementary Material-

Yufei Shi^{1,5†}, Weilong Yan^{2†}, Gang Xu⁴, Yumeng Li³, Yucheng Chen^{1,5},
Zhenxi Li^{1,5}, Fei Richard Yu⁴, Ming Li⁴, Si Yong Yeo^{1,5}

¹MedVisAI Lab ²National University of Singapore ³Nankai University

⁴Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)

⁵Lee Kong Chian School of Medicine, Nanyang Technological University

In this supplemental material, the readers can find:

- Experimental details about the ablation study of token number;
- Details about multiple metrics utilized for personalized QA quality;
- More examples of recognition question answering;
- More experiment detail about Multi Character training
- Some examples of existential questions and answers for single and two entities;
- Presentation of templates for GPT prompt queries;
- Presentation of 25 characters;
- Some examples of Similar Negative Sample
- Some examples of Movie Datasets and Vlog Datasets
- Our code is shown in the PVChat code zip files

1. Ablation Study About the Token Number

We set each question prompt corresponding to $N = 16$ tokens per character. If $N = 0$, training is only with the `<sks>` token. As the token number increases, the metrics overall increase. However, when the number is larger than 16, the performance decreases. This is because too many tokens can make the model hard to capture the characteristics of the subject. The detailed experimental results are shown in Tab.1.

Number	Acc \uparrow	BLEU \uparrow	BS \uparrow	ES \uparrow	DC \uparrow
0	0.801	0.495	0.939	5.00	4.00
4	0.871	0.592	0.951	5.00	4.50
8	0.890	0.553	0.949	5.00	4.58
12	0.895	0.564	0.949	5.00	4.41
16	0.922	0.606	0.952	4.74	4.38
20	0.882	0.554	0.949	4.96	4.25

Table 1. Ablations on the number of tokens.

2. Details of the Evaluation Metrics

Here we add more details of the metrics which are utilized to measure the quality of personalized video chat:

(1) Accuracy: Specifically designed to evaluate binary existence questions, measuring the model’s ability to correctly identify the presence or absence of objects.

(2) BLEU and BERTScore: These metrics quantify the textual similarity between generated responses and ground truth answers, capturing linguistic precision and semantic alignment.

(3) Entity Specificity (ES): Evaluated on a 1-5 scale, this metric assesses whether responses contain personalized, contextually relevant details rather than generic statements.

(4) Descriptive Completeness (DC): Also rated on a 1-5 scale, DC measures the logical coherence, factual correctness, and comprehensive nature of responses. This metric evaluates whether answers are thoroughly developed with appropriate supporting details and proper reasoning.

3. More Examples of Our PVChat Model

From Fig.1 to Fig.5, more personalized video chat examples of various scenarios are displayed. For example, in Fig.1, there are three characters named `<Cl>`, `<Ja>` and `<Xo>` in a laboratory scene, which is quite challenging for video understanding. Our PVChat not only accurately identifies these three men by their names but also gives reasonable suggestions for some additional questions about giving gifts. What’s more, when the model is asked about their behavior, it successfully captures their locations and can do reasoning to guess this is a professional setting. This demonstrates that our model has strong personalized reasoning capability, even in a multiple subject setting. More details can be seen in the figures.



More Examples of PVChat	
Ref Video	 <div style="display: flex; justify-content: space-around; margin-top: 5px;"> <CI> <Ja> <Xo> </div>
Query Video	
Question	Can you detect <CI>, <Xo>, or <Ja> in this video clip?
PVChat (ours)	I've detected <CI>, <Xo>, and <Ja> in the video.
Question	If I want to buy three gift for <CI>, <Xo> and <Ja>, can you give me some idea?
PVChat (ours)	<CI>, <Xo>, and <Ja> might be interested in books, movies, or video games.
Question	If I want to write a ins style poster about <CI>, <Xo> and <Ja> doing today, can you give me a detailed example?
PVChat (ours)	Certainly! Here's an example of a <CI>, <Xo>, and <Ja> doing something together: <CI>, <Xo>, and <Ja> are standing in a room with a painting on the wall behind them. <Xo> is wearing a white coat. <CI> is looking at <Xo> and smiling. <Ja> is wearing a black shirt and is looking at <CI>. <CI> is wearing a white coat and is looking at <Ja>. All three of them are standing next to each other and appear to be in a professional setting.

Figure 1. Example of PVChat.

4. More experiment detail about Multi-Character training

Epoch One person training needs 8 epochs, and two people training need 16 epochs, and three people training need 24 epochs.

ReMoH Within the ReMoH architecture, we implement cross-attention mechanisms at alternating layers to facilitate interaction between video embeddings (extracted by the visual encoder) and query embeddings. Our approach adap-

tively scales with the number of subjects: for single-person training, we incorporate an additional $16 \times \text{token_length} \times 768$ dimensional feature space; for two-person training, this expands to $32 \times \text{token_length} \times 768$ and for three-person training, it further increases to $48 \times \text{token_length} \times 768$.

Crucially, these query tokens are associated with corresponding masks that are only deactivated during the training of their respective subject videos. This selective masking strategy ensures that each query embedding specifically attends to its designated subject, effectively preventing cross-


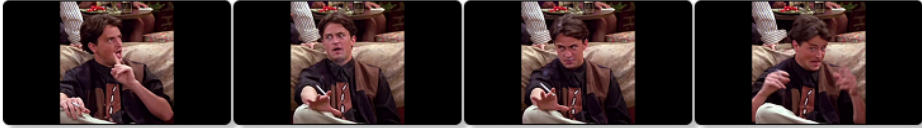
More Examples of PVChat	
Ref Video <Ch>	
Query Video	
Question	Can you describe <Ch> 's behavior in this sequence?
PVChat (ours)	<p><Ch> in the video is seen sitting on a couch and holding a cigarette.</p> <p><Ch> is making various hand gestures and facial expressions, including raising <Ch>'s hand, making a peace sign, and putting <Ch>'s finger to <Ch>'s mouth.</p>

Figure 2. Example of PVChat.



More Examples of PVChat	
Ref Video <Ra>	
Query Video	
Question	Can you describe <Ra> 's behavior in this sequence?
PVChat (ours)	<p>The <Ra> in the video is seen making facial expressions and gestures that suggest <Ra> are in a conversation or discussion. <Ra> are wearing a black shirt and are standing in front of a brick wall.</p>

Figure 3. Example of PVChat.

contamination between different identity representations.

Special tokens For multi-person scenarios, we employ the special tokens [PERSON] and [\PERSON] as delimiters to

clearly delineate the boundary between one person’s SKS tokens and another’s. This demarcation approach enables the model to accurately distinguish when the sequence of tokens



More Examples of PVChat	
Ref Video <Ho>	
Query Video	
Question	Can you carefully determine if <Ho> or <Sh> exists in this video?
PVChat (ours)	The video contains <Ho> , but there's no sign of <Sh> .

Figure 4. Example of PVChat.


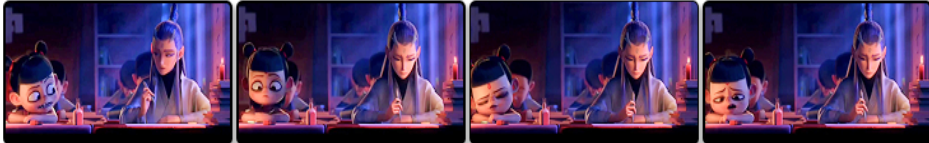
More Examples of PVChat	
Ref Video <Nz> <Ab>	
Query Video	
Question	Can you describe the activity that <Nz> and <Ab> are engaged in if you could find them in the video?
PVChat (ours)	<Nz> and <Ab> in the video are engaged in writing on a piece of paper.

Figure 5. Example of PVChat.

associated with one individual ends and those belonging to another begin.

5. Example of 4 Different Questions

From Tab 2 to Tab 5, the detailed existence question for one entity and two entities are displayed. We show the other

three questions about the appearance, action, and location in Tab.6.

Index	Question	Yes Answer
1	Is there any trace of <sks> in this footage?	Yes, <sks> is in this video.
2	Can you detect <sks> in this video clip?	I can confirm that <sks> appears.
3	Does <sks> show up anywhere in this recording?	<sks> is present in this recording.
4	Is <sks> visible in this video?	The video contains <sks>.
5	Could you verify if <sks> is here?	I've identified <sks>.
6	Does this footage include <sks>?	<sks> is shown in this video.
7	Can you spot <sks> in this clip?	Yes, <sks> appears here.
8	Is <sks> present in this video?	I can verify that <sks> is present.
9	Does <sks> appear in this footage?	The footage shows <sks>.
10	Can you tell if <sks> is shown here?	<sks> is in this video clip.
11	Is <sks> in this video segment?	I've detected <sks>.
12	Can you confirm <sks>'s presence?	Yes, <sks> is featured.
13	Does this clip contain <sks>?	The video includes <sks>.
14	Is <sks> featured in this recording?	I can see <sks>.
15	Can you find <sks> in this video?	<sks> is definitely here.
16	Is <sks> shown in any frame?	Yes, I've found <sks>.
17	Does this video show <sks>?	This video shows <sks>.
18	Is <sks> visible anywhere?	<sks> is visible.
19	Can you see <sks>?	Yes, <sks> has been captured.
20	Is <sks> in this video?	The video clearly shows <sks>.
21	Can you recognize <sks>?	I've spotted <sks>.
22	Does <sks> appear at all?	<sks> appears in this video.
23	Is <sks> recorded here?	Yes, this footage contains <sks>.
24	Can you identify <sks>?	I can recognize <sks>.
25	Is <sks> present?	<sks> is clearly visible.

Table 2. Example of positive existence verification answering (Single entity).

Index	Question	No Answer
1	Is there any trace of <sks> in this footage?	No, <sks> is not in this video.
2	Can you detect <sks> in this video clip?	I cannot detect <sks>.
3	Does <sks> show up anywhere in this recording?	This video does not contain <sks>.
4	Is <sks> visible in this video?	<sks> is not shown.
5	Could you verify if <sks> is here?	There is no sign of <sks>.
6	Does this footage include <sks>?	<sks> does not appear.
7	Can you spot <sks> in this clip?	I can confirm <sks> is not here.
8	Is <sks> present in this video?	The footage does not include <sks>.
9	Does <sks> appear in this footage?	There's no evidence of <sks>.
10	Can you tell if <sks> is shown here?	<sks> is not in this video.
11	Is <sks> in this video segment?	I've checked, <sks> is not present.
12	Can you confirm <sks>'s presence?	This video does not show <sks>.
13	Does this clip contain <sks>?	I see no sign of <sks>.
14	Is <sks> featured in this recording?	<sks> is absent.
15	Can you find <sks> in this video?	The video does not show <sks>.
16	Is <sks> shown in any frame?	I cannot find <sks>.
17	Does this video show <sks>?	<sks> is not visible.
18	Is <sks> visible anywhere?	I can verify <sks> is not here.
19	Can you see <sks>?	The video has no <sks>.
20	Is <sks> in this video?	<sks> does not exist in this video.
21	Can you recognize <sks>?	I find no trace of <sks>.
22	Does <sks> appear at all?	This clip does not contain <sks>.
23	Is <sks> recorded here?	<sks> is not present.
24	Can you identify <sks>?	I cannot identify <sks>.
25	Is <sks> present?	There is no <sks> here.

Table 3. Example of negative existence verification answering (Single entity).

Index	Question	Both Yes Answer
1	Is there any trace of <sks1> or <sks2> in this footage?	Both <sks1> and <sks2> are present in this video.
2	Can you detect <sks1> or <sks2> in this video clip?	I can detect both <sks1> and <sks2> in the footage.
3	Do <sks1> or <sks2> show up anywhere in this recording?	The video shows both <sks1> and <sks2> clearly.
4	Are <sks1> or <sks2> visible in this video?	<sks1> and <sks2> are both visible in this recording.
5	Could you verify if <sks1> or <sks2> are here?	I've identified both <sks1> and <sks2> in the clip.
6	Does this footage include <sks1> or <sks2>?	Both <sks1> and <sks2> appear in this video.
7	Can you spot <sks1> or <sks2> in this clip?	The footage contains both <sks1> and <sks2>.
8	Are <sks1> or <sks2> present in this video?	I can see both <sks1> and <sks2> in frame.
9	Do <sks1> or <sks2> appear in this footage?	<sks1> and <sks2> are both featured in this video.
10	Can you tell if <sks1> or <sks2> are shown here?	The recording shows both <sks1> and <sks2> present.
11	Are <sks1> or <sks2> in this video segment?	I've spotted both <sks1> and <sks2> in the footage.
12	Can you confirm <sks1> or <sks2>'s presence?	Both <sks1> and <sks2> are captured in this clip.
13	Does this clip contain <sks1> or <sks2>?	The video includes both <sks1> and <sks2>.
14	Are <sks1> or <sks2> featured in this recording?	I can confirm the presence of both <sks1> and <sks2>.
15	Can you find <sks1> or <sks2> in this video?	<sks1> and <sks2> are both shown in the recording.
16	Are <sks1> or <sks2> shown in any frame?	Both figures, <sks1> and <sks2>, are visible.
17	Does this video show <sks1> or <sks2>?	I've found both <sks1> and <sks2> in the video.
18	Are <sks1> or <sks2> visible anywhere?	The footage displays both <sks1> and <sks2>.
19	Can you see <sks1> or <sks2>?	Both <sks1> and <sks2> are identifiable here.
20	Are <sks1> or <sks2> in this video?	I can recognize both <sks1> and <sks2>.
21	Can you recognize <sks1> or <sks2>?	<sks1> and <sks2> both appear in this recording.
22	Do <sks1> or <sks2> appear at all?	The video features both <sks1> and <sks2>.
23	Are <sks1> or <sks2> recorded here?	Both <sks1> and <sks2> are clearly visible.
24	Can you identify <sks1> or <sks2>?	I've detected the presence of both <sks1> and <sks2>.
25	Are <sks1> or <sks2> present?	The clip shows both <sks1> and <sks2>.

Table 4. Example of positive existence verification question answering (Two entities).

Index	sks1 Yes, sks2 No	Both No
1	Is there any trace of <sks1> or <sks2> in this footage?	I can confirm that <sks1> appears, but <sks2> is not present. Neither <sks1> nor <sks2> appear in this video.
2	Can you detect <sks1> or <sks2> in this video clip?	The video shows <sks1>, though there's no sign of <sks2>. I cannot detect either <sks1> or <sks2>.
3	Do <sks1> or <sks2> show up anywhere in this recording?	<sks1> is visible, but <sks2> is absent. The video contains neither <sks1> nor <sks2>.
4	Are <sks1> or <sks2> visible in this video?	I've detected <sks1>, while <sks2> does not appear. Both <sks1> and <sks2> are absent.
5	Could you verify if <sks1> or <sks2> are here?	The video contains <sks1>, but <sks2> is not shown. There is no sign of either <sks1> or <sks2>.
6	Does this footage include <sks1> or <sks2>?	<sks1> is present, however <sks2> is not in this clip. Neither <sks1> nor <sks2> are shown.
7	Can you spot <sks1> or <sks2> in this clip?	I can see <sks1>, but there's no trace of <sks2>. I confirm both <sks1> and <sks2> are not present.
8	Are <sks1> or <sks2> present in this video?	The footage includes <sks1>, though <sks2> is not visible. The footage does not include <sks1> or <sks2>.
9	Do <sks1> or <sks2> appear in this footage?	<sks1> appears, but <sks2> is not featured. There's no evidence of either <sks1> or <sks2>.
10	Can you tell if <sks1> or <sks2> are shown here?	I've spotted <sks1>, while <sks2> is nowhere to be seen. Neither <sks1> nor <sks2> are visible.
11	Are <sks1> or <sks2> in this video segment?	<sks1> is clearly visible, but <sks2> is not. I've checked, both <sks1> and <sks2> are absent.
12	Can you confirm <sks1> or <sks2>'s presence?	The recording shows <sks1>, though <sks2> is absent. This video shows neither <sks1> nor <sks2>.
13	Does this clip contain <sks1> or <sks2>?	I can identify <sks1>, but <sks2> doesn't appear. I see no sign of <sks1> or <sks2>.
14	Are <sks1> or <sks2> featured in this recording?	<sks1> is present, while <sks2> is not. Both <sks1> and <sks2> are not in the recording.
15	Can you find <sks1> or <sks2> in this video?	The clip features <sks1>, but there's no sign of <sks2>. The video does not contain <sks1> or <sks2>.

Table 5. Example of mixed existence verification question answering (Two entities).

Index	Question
1	What activity is <sks> engaged in during this video?
2	Could you describe what <sks> is doing in this footage?
3	What specific actions can you observe <sks> performing in this recording?
4	What movements or actions does <sks> perform here?
5	Can you describe <sks>'s behavior in this sequence?
6	What is <sks> wearing in this video?
7	Could you describe <sks>'s outfit in this footage?
8	What color and style of clothing is <sks> dressed in?
9	How would you describe <sks>'s appearance and attire?
10	What notable features can you see in <sks>'s clothing?
11	Where is <sks> positioned in this video?
12	What color and style of clothing is <sks> dressed in?
13	Can you describe <sks>'s location relative to others?
14	Which part of the scene does <sks> appear in?
15	How does <sks>'s position change throughout the video?
16	Where can <sks> be found in this footage?

Table 6. Example of negative existence verification question answering (Single entity).

6. Example of queries prompt for GPT or Internvideo

As shown in 6, this prompt will be used to get the age and gender of the character in the video by the video understanding model, and the age and gender information will be used in the following steps. And for the general answer from the video, understanding model will use the second one prompt to query ChatGPT to capture the general description like the human, he, she, and this will be replaced by the specific <sks> and then to construct our personalized QA pairs.

As shown in 7, leveraging the character's age and gender attributes extracted in the previous stage, we enrich these core identity markers with detailed descriptors to create comprehensive prompts for (image+text)-to-image models. This approach significantly enhances identity consistency across generated images, maintaining coherent character representation while accommodating diverse poses and contexts. The resulting outputs demonstrate superior identity preservation compared to methods lacking such demographic anchoring.

7. Presentation of 25 characters

As shown in Fig. 8, our dataset is composed of Friends(6), Good Doctor(5), Ne Zha(2), doctor(3), patient(3), Big Bang(6)) 25 characters.

8. Example of similar negative Samples

As shown in Fig.9, We display some similar Negative Samples.

9. Examples in Movie Datasets and Vlog video

We show some sample about Movie Datasets in Fig.10 and Vlog datasets in Fig. 11.

Video Caption to Get Age and Gender by Internvideo2.

According to the video and just told me two words about gender from "male, female" and age from "young, middle-aged, elderly" about the protagonist of the video.

[Video Input]

Caption Replace by ChatGPT-4o.

You are a helpful assistant. Help me replace all the descriptions of people in the following paragraph, such as human, he, she, etc., with <person>. Just return the modified one:

[Input caption]

Figure 6. Prompt for Internvideo and GPT query

Personalized Image Generation Prompt by Photomaker.

{age} {gender} img sitting at a clean desk in a modern office, facing camera directly, wearing business suit, professional lighting, clear facial details, 8k quality, photorealistic.

{age} {gender} img in photography studio, plain white background, standing centered, direct eye contact, perfect studio lighting, professional portrait, ultra HD, photorealistic.

{age} {gender} img sitting alone at wooden table in coffee shop, facing straight to camera, soft natural lighting, wearing casual outfit, blurred cafe background, crystal clear facial features.

{age} {gender} img in minimal living room, sitting on plain sofa, looking straight at camera, warm indoor lighting, wearing comfortable clothes, high resolution portrait.

{age} {gender} img in modern office space, sitting at minimalist desk, wearing professional attire, making eye contact with camera, clean background, natural lighting from window.

{age} {gender} img in home library, sitting in armchair with book, facing forward, warm lighting, simple background with bookshelves, detailed face capture, casual smart clothing.

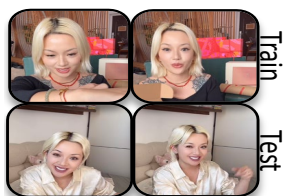
Figure 7. Prompt for Photomaker synthetic the ID consis photo



Figure 8. Display of all dataset.



Figure 9. Similar Negative Samples.



Q: Can you describe **<Ye>**'s behavior in this sequence?
 PVChat: The **<Ye>** in the video is seen smiling and looking directly at the camera. **<Ye>** is wearing a white shirt and has short blonde hair.

Figure 10. Movie Datasets Results



Q: Can you explain what this video is showing and tell me about the character emotion?
 PVChat: The **<Hermione>** is pointing a wand-like object towards the viewer in a dimly lit room. The character appears to be in a state of concentration, with a serious expression on their face.
 AutoAD_II MAC datasets: She looks at them sadly, and raises her wand.

Figure 11. Vlog Sample