VueICL: Entity-Aware Video Question Answering through In-Context Learning of Visual Annotations

Shahaf Levinson [1] **Ram Elgov** [1] **Yonatan Benizri** [1] **Idan Schwartz** [2]

Abstract

Large multi-modal models (LMMs) integrate various modalities such as text, images, and, more recently, long videos. Despite significant advancements in long video understanding, spatio-temporal reasoning, and question answering, these models struggle to recognize specific subjects within videos and to answer entity-aware questions. Current methods addressing similar limitations in vision-language models (VLMs) rely on fine-tuning and few-shot learning to introduce specific objects or individuals present in images. However, these approaches are largely confined to image retrieval and captioning tasks. We find these methods impractical for entity-aware video understanding due to the substantial visual content in videos and the additional spatio-temporal overhead, which incurs significant computational costs.

While in-context learning (ICL) with videos and images of entities has demonstrated competitive performance in vision-language tasks within a training-free environment, applying similar methods to long videos faces challenges due to limited context length. To overcome these challenges, we propose Video Understanding Entities In-Context Learning (VueICL), a simple yet effective video in-context learning framework and benchmark for answering entity-aware questions about videos. Our approach introduces a two-stage pipeline: in the first stage, we add visual annotations to each video, and in the second stage, we use these annotations to assist the inference process.

Experimental results on our new benchmark demonstrate performance gains compared to existing baselines using state-of-the-art video LMMs, opening up significant opportunities for the efficient personalization of long video understanding tasks.

The code for our project is available at: https://github.com/ram-elgov/VueICL

[*]Equal contribution [1]Tel Aviv University [2]Bar-Ilan University, Ramat Gan, Israel. Correspondence to: Shahaf Levinson <shahafl@mail.tau.ac.il>, Ram Elgov <ramelgov@mail.tau.ac.il>, Yonatan Benizri <benizrilevi@mail.tau.ac.il>, Idan Schwartz <idan.schwartz@biu.ac.il>.

1. Introduction

Understanding video content through natural language queries has become a critical challenge in multimodal artificial intelligence, particularly with the advent of large language models integrated with vision capabilities (VLMs). While significant progress has been made in enabling these models to address general queries about visual and multimodal data, achieving user-specific personalization—tailoring model responses based on user-defined concepts or preferences—remains an ongoing challenge.

In the realm of image-based personalization, previous research has shown that personalization can be achieved through fine-tuning models or by training concept embeddings (Alaluf et al., 2024). These methods allow models to identify user-specific objects, animals, or other entities by mapping them into a shared multimodal representation space. However, extending these techniques to video content introduces additional complexities, such as the need to process temporal information and adapt to dynamic changes in scenes, particularly for person-specific personalization tasks.

Personalization using annotation: In this work, we propose a novel approach for enabling user-specific queries about video content, specifically focusing on the personalization of individuals in videos. Unlike prior methods that require extensive model retraining or embedding fine-tuning, our solution leverages annotations on video frames in conjunction with state-of-the-art VLMs. By annotating video frames with red bounding boxes around individuals and providing their corresponding names, we explicitly introduce a personalized context.

We use prompts such as: "`The red box represents a person's name. [Question]`" to instruct the VLMs to associate the red boxes with specific individuals.

This simple yet effective method allows the model to incorporate person-specific context and answer user-specific queries about the annotated video, such as:

- Identifying properties regarding specific individuals ("What color is 1's shirt?")
- Identifying actions performed by a specific individual ("Who is speaking on the phone in the video?")
- Providing personalized video captions ("What does 1 do in the video?")

Our approach avoids the computational cost of model fine-tuning or embedding training, instead utilizing the powerful generalization capabilities of state-of-the-art LMMs. Through this method, we bridge the gap between generic video understanding and personalized video interaction, providing a scalable solution for applications such as personalized video search, content analysis, and interactive multimedia systems.

2. Related Work

LongVU LongVU (Shen et al., 2024) is a spatiotemporal adaptive compression mechanism designed to improve the capacity of Multimodal Large Language Models (MLLMs) to process and comprehend long video content. This project tackles the challenge of limited context size in MLLMs when handling extensive video data. The LongVU architecture includes:

Temporal Reduction: Utilizes DINOv2 (Oquab et al., 2024) features to detect and eliminate highly similar, redundant frames, reducing temporal redundancy. Afterwards, the model fuses the remaining frame features from both SigLIP (Zhai et al., 2023) and DINOv2.

Selective Feature Reduction: Employs text-guided cross-modal queries to selectively condense frame features, retaining critical visual information.

Spatial Token Reduction: Compresses spatial tokens across frames by leveraging their temporal dependencies.

Adaptive Compression: Balances the ability to process numerous frames within a fixed context length while minimizing loss of visual details.

Our tests revealed that LongVU outperformed other models in delivering detailed descriptions of video content, excelling at identifying scenes, characters, and interactions in response to user queries.

MyVLM The paper (Alaluf et al., 2024) presents an innovative method for tailoring large-scale vision-language models (VLMs) to user-specific concepts. This approach enables VLMs to learn and reason about user-defined elements, such as identifying individuals in images and describing their actions. Key features of MyVLM include:

External Concept Heads: These act as toggles, enhancing VLMs to detect specific target concepts in images.

Concept Embedding Learning: MyVLM introduces a new concept embedding within the VLM's intermediate feature space, guiding the language model to incorporate the target concept seamlessly into its responses.

Application to VLM Models: The method is applied to BLIP-2 (Li et al., 2023) and LLaVA (Liu et al., 2023) models, demonstrating its effectiveness in tasks like personalized image captioning and visual question-answering.

The researchers showcase MyVLM's ability to personalize concepts, including specific objects and individuals, using only a few example images. The approach generalizes well to unseen images of the learned concepts while maintaining the model's original performance on unrelated inputs.

Meta-Personalizing Vision-Language Models to Find Named Instances in Video In their work, (Yeh et al., 2023), the authors tackle the problem of personalized video searches, such as identifying moments featuring specific objects or instances (e.g., "My dog Biscuit"). They propose an approach that extends the VLM's token vocabulary with instance-specific embeddings, combining shared category features with learned instance-specific representations. Personalization is achieved without explicit human supervision by leveraging the automatic identification of named visual instances in videos using transcripts and vision-language embedding similarity. To benchmark progress in this domain, the authors introduce *This-Is-My*, a dataset designed for personal video instance retrieval.

3. Preliminaries

Vision-Language Models (VLMs) are multimodal models that aim to bridge the gap between visual and textual modalities. These models are designed to process and integrate information from both images and text, enabling applications such as image captioning, visual question answering, and cross-modal retrieval. Notable examples include CLIP (Radford et al., 2021), which aligns text and image embeddings for robust zero-shot classification, BLIP-2 (Li et al., 2023), and LLaVA (Liu et al., 2023), which enhance this framework by integrating vision-language pre-training with caption generation and question-answering capabilities.

Vision-Language capabilities have nowadays been integrated into many state-of-the-art models, reflecting the growing importance of multimodal understanding. Models like GPT-4 (OpenAI, 2023) and Gemini (Team, 2024) have extended their functionalities to incorporate VLM capabilities,

enabling seamless interaction across textual and visual inputs.

Face Recognition is a critical component in many applications, such as security, surveillance, and user authentication. The face recognition task has been extensively studied through the lens of deep neural networks. One of the state-of-the-art approaches presented in (**?**) involves the Residual Network (ResNet), which addresses the degradation problem in deep networks by reformulating layers to learn residual functions with reference to layer inputs. The authors show that this technique enables the successful training of significantly deeper networks compared to earlier models, with depths of up to 152 layers, achieving state-of-the-art performance on the ImageNet dataset.

Leveraging these advancements in deep learning, tools like the dlib (King, 2009) library have gained popularity for their robustness and efficiency in implementing these architectures. The `face_recognition` (Geitgey, 2017) library, which builds upon dlib, enables lightweight and easy-to-use APIs for face recognition, allowing users to solve face recognition tasks without requiring an in-depth understanding of the underlying architecture.

In-Context Learning (ICL) (Dong et al., 2024) is a paradigm where models learn from a few examples provided in the context. Many studies have shown that large language models (LLMs) can perform a series of complex tasks through ICL, such as solving mathematical reasoning problems (**?**). These strong abilities have been widely verified as emergent abilities for large language models (**?**). The key idea of in-context learning is to learn from analogy. First, ICL requires a few demonstration examples to form a prompt context. These examples are usually written in natural language templates. Then, ICL concatenates a query question and the piece of prompt context together to form the input, which is then fed into the language model for prediction. Unlike supervised learning, which requires a training stage that uses backward gradients to update model parameters, ICL does not perform parameter updates. The model is expected to learn the pattern hidden in the demonstrations and accordingly make the right prediction. One of the greatest advantages of ICL is that it is a training-free learning framework, which greatly reduces the computational costs for adapting the model to new tasks.

4. Method

4.1. In-Context Learning (ICL)

For the in-context learning (ICL) approach, we provided the vision-language model (VLM) with a set of image files, each associated with a specific label (a person's name) through a descriptive prompt that links the file name to the corresponding individual. This process effectively added a mapping between a person's name and their images into the VLM's context. Once this association was established, we attached the video along with the accompanying question and sent them to the VLM for a response. This method leverages the added context to enable the VLM to better understand and answer personalized questions about the video.

4.2. Annotation

Our pipeline consists of two main components: face annotation using face recognition and multimodal VLMs.

Face Annotation: We first annotate video frames with red bounding boxes around individuals and provide their corresponding names. We used the open-source Python library `face_recognition` (Geitgey, 2017), which is built using dlib (King, 2009), to detect faces in the video frames and draw red bounding boxes around them. We then manually labeled the individuals in the video frames and provided corresponding textual labels (e.g., 1, 2, etc.).

Multimodal VLMs: We use multimodal VLMs to process the annotated video frames and answer user-specific queries. Empirically, we found that the best model for our task is the LongVU model (Shen et al., 2024) (among the ones we tried are VideoLLAMA, Gemini, and ChatGPT). After choosing our model, we empirically examined different prefixes to instruct the model to associate the red boxes with specific individuals. We found that the best prefixes were:

- **Short prefix:** "The red box shows a person's name. [Question]"
- **Long prefix:** "You are a multimodal understanding model. The input video contains scenes where characters are annotated with red bounding boxes labeled as 1, 2, etc. Your task is to answer questions based on the visual content of the video. Use the red bounding boxes and their labels to determine which character performs the described action or matches the described attributes.

  Focus on:

  - The visual content inside the red bounding boxes. - The labels (1, 2, etc.) to identify characters. - Attributes or actions described in the question.

  Question: [Question]"

Some semantically similar prefixes were also tried but did not perform well:

- "The red box indicates a person's name."
- "Each red box represents a person's name."

5. Experiments

5.1. Dataset

Since there is no public dataset for the task of entity-aware video understanding, we provide a new benchmark curated for this task using a semi-automatic method. The dataset consists of 22 raw videos, each containing at least two individuals. The selected videos depict interactions between individuals, allowing us to explore questions about their appearance, interactions, and activities throughout the video. We manually annotated the videos with red bounding boxes around the individuals and provided their corresponding names.

5.2. Evaluation Metrics

We evaluate the different methods using a set of closed-ended questions about the individuals in the videos, such as:

- What color is 1's shirt?
- Who is speaking on the phone in the video?
- What does 1 do in the video?

These closed questions are designed to test the model's ability to identify properties, actions, and interactions of specific individuals in the video content, providing a comprehensive evaluation of the model's personalized video understanding capabilities.

5.3. Baselines

While existing video benchmarks such as ConCon-Chi (Rosasco et al., 2024) and STAR (Wu et al., 2021) provide comprehensive evaluation frameworks for various vision-language tasks, there is no curated benchmark specifically designed for entity-aware question answering in videos. To address this gap, we established three baselines to evaluate the effectiveness of our proposed method:

1. **Empty Video**: This baseline involves testing prompts against an empty video to assess the model's performance in the absence of any visual content. It serves as a control to determine the baseline level of question-answering accuracy without any visual input.

2. **No Annotation Long Prompt**: In this baseline, we test prompts against the original video without any annotations. The purpose is to evaluate whether the model exhibits a bias toward assigning specific labels to individuals or if it can inherently recognize and reference entities without explicit visual cues.

Table 1. Results of our different methods over the evaluation metrics. **Questions** refer to the number of questions solved, and **Videos** to the number of videos where all questions were solved correctly.

METHOD	QUESTIONS	VIDEOS
LONGVU EMPTY VIDEO	44/100	1/22
LONGVU NO ANNOTATION	54/100	3/22
QWEN2-VL	49/100	4/22
VUEICL LONG PROMPT	**67/100**	**6/22**
VUEICL SHORT PROMPT	**68/100**	**6/22**

3. **Qwen2-VL (Wang et al., 2024) with Vision IDs**: Utilizing the Qwen2-VL model as a state-of-the-art (SOTA) vision-language model baseline allows us to benchmark our method against a robust existing model. We enhanced Qwen2-VL's capability in handling multiple visual inputs by incorporating unique textual identifiers for each image and video, enabling the model to reference and differentiate between multiple entities within a single conversation context.

These baselines provide critical insights into the model's behavior and its dependency on visual input. By comparing our proposed **VueICL** framework against these baselines, we can comprehensively assess the improvements in entity-aware video question answering and the effectiveness of incorporating visual annotations without the need for extensive model fine-tuning.

These baselines provide critical insights into the model's behavior and its dependency on visual input. By comparing our proposed **VueICL** framework against these baselines, we can comprehensively assess the improvements in entity-aware video question answering and the effectiveness of incorporating visual annotations without the need for extensive model fine-tuning.

5.4. Results

Table 1 presents the performance of our proposed methods compared to established baselines on the entity-aware video understanding benchmark. The Empty Video baseline, which involves testing prompts against an empty video, achieved 44 out of 100 questions answered correctly and 1 out of 22 videos where all questions were correctly answered. Introducing annotations without additional prompts through the LongVU No Annotation method improved performance to 54/100 questions and 3/22 videos. The Qwen2-VL model (Wang et al., 2024), a state-of-the-art vision-language model, achieved 49/100 questions and 4/22 videos, outperforming the basic annotation baseline but still lagging behind our proposed methods. Our VueICL framework, uti-

lizing both long and short prompts, significantly enhanced performance, achieving 67/100 and 68/100 questions correctly answered, and 6/22 videos with all questions correctly answered for the long and short prompts, respectively. These results demonstrate that VueICL effectively leverages in-context learning of visual annotations to surpass existing baselines and state-of-the-art models, highlighting its potential for personalized and efficient video question answering. The marginal improvement of the short prompt over the long prompt underscores the effectiveness of concise instructional cues in facilitating model comprehension and response accuracy.

6. Limitations

Our current approach does not incorporate **audio** information, which may limit the ability to fully capture the context of interactions within videos. The performance of our system also depends heavily on the **capabilities of the VLM** and the **effectiveness of the face recognition** algorithm in accurately identifying individuals within the video frames. Challenges such as occlusions, low-resolution images, or rapid movements may affect the accuracy of person detection and, consequently, the reliability of the analysis. Additionally, our method relies on manual annotation of video frames, which can be time-consuming and may not scale effectively to larger datasets. Addressing these limitations in future iterations could enhance the overall performance and scalability of the system.

7. Conclusions

In the same way that Object-Oriented Programming (OOP) revolutionized software design by enabling the referencing of specific instances of general concepts, we anticipate that large multi-modal models (LMMs) will similarly advance through entity-aware reasoning. This paper introduces Video Understanding Entities In-Context Learning (VueICL), a novel framework that leverages in-context learning of visual annotations to enhance the ability of state-of-the-art LMMs in answering entity-aware questions about long videos. By employing a two-stage pipeline that first annotates video frames with visual markers and corresponding entity labels, and then integrates these annotations into the inference process, VueICL effectively bridges the gap between generic video understanding and personalized video interaction without the computational overhead of traditional fine-tuning methods. Experimental results on our newly curated benchmark demonstrate significant performance gains over existing baselines, including the robust Qwen2-VL model, highlighting the efficacy of our annotation-based approach. Despite its promising advancements, VueICL currently does not incorporate audio information and relies on the accuracy of face recognition algorithms, which may limit its effectiveness in scenarios with occlusions or low-resolution footage. Future work will focus on integrating multimodal inputs beyond visual data and enhancing entity detection robustness to address these challenges. Overall, VueICL represents a significant step toward more intelligent and personalized multi-modal AI systems, paving the way for advanced applications in personalized video search, content analysis, and interactive multimedia technologies.

References

Alaluf, Y., Richardson, E., Tulyakov, S., Aberman, K., and Cohen-Or, D. Myvlm: Personalizing vlms for user-specific queries. https://arxiv.org/abs/2403.14599, 2024. URL https://arxiv.org/abs/2403.14599.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., Chang, B., Sun, X., Li, L., and Sui, Z. A survey on in-context learning. https://arxiv.org/abs/2301.00234, 2024. URL https://arxiv.org/abs/2301.00234.

Geitgey, A. Face recognition. https://github.com/ageitgey/face_recognition, 2017. URL https://github.com/ageitgey/face_recognition. Accessed: 2024-12-11.

King, D. E. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. URL https://www.jmlr.org/papers/v10/king09a.html. Available at https://www.jmlr.org/papers/v10/king09a.html.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. https://arxiv.org/abs/2301.12597, 2023. URL https://arxiv.org/abs/2301.12597.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. https://arxiv.org/abs/2304.08485, 2023. URL https://arxiv.org/abs/2304.08485.

OpenAI. Chatgpt: Language model for conversational ai. https://openai.com/index/chatgpt-can-now-see-hear-and-speak/, 2023. URL https://openai.com/index/chatgpt-can-now-see-hear-and-speak/. Accessed: 2024-12-11.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma,

V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision. https://arxiv.org/abs/2304.07193, 2024. URL https://arxiv.org/abs/2304.07193.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. https://arxiv.org/abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Rosasco, A., Berti, S., Pasquale, G., Malafronte, D., Sato, S., Segawa, H., Inada, T., and Natale, L. ConCon-Chi: Concept-Context Chimera Benchmark for Personalized Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., Liu, Z., Xu, H., Kim, H. J., Soran, B., Krishnamoorthi, R., Elhoseiny, M., and Chandra, V. Longvu: Spatiotemporal adaptive compression for long video-language understanding. https://arxiv.org/abs/2410.17434, 2024. URL https://arxiv.org/abs/2410.17434.

Team, G. Gemini: A family of highly capable multimodal models. https://arxiv.org/abs/2312.11805, 2024. URL https://arxiv.org/abs/2312.11805.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. https://arxiv.org/abs/2409.12191, 2024. URL https://arxiv.org/abs/2409.12191.

Wu, B., Yu, S., Chen, Zhenfang, T. J. B., and Gan, C. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Yeh, C.-H., Russell, B., Sivic, J., Heilbron, F. C., and Jenni, S. Meta-personalizing vision-language models to find named instances in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19123–19132, 2023.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. https://arxiv.org/abs/2303.15343, 2023. URL https://arxiv.org/abs/2303.15343.