THAVQA: a German task-oriented VQA dataset annotated with human visual attention

Moritz Kronberger[1,†], Viviana Ventura[1,*,†]

[1]Technische Hochschule Augsburg, An der Hochschule 1, 86161 Augsburg, Germany

Abstract

Video question answering (VQA) is a challenging task that requires models to generate answers by using both information from text and video. We present Task-oriented Human Attention Video Question Answering (THAVQA), a new VQA dataset consisting of third- and first- person videos of an instructor using a sewing machine. The sewing task is formalized step-by-step in a script: each step consists of a video annotated with German language open-ended question and answer (QA) pairs and with human visual attention. The paper also includes a first assessment of the performance of a pre-trained Multimodal Large Language Model (MLLM) in generating answers to the questions of our dataset across different experimental settings. Results show that our task-oriented dataset is challenging for pre-trained models. Specifically, the model struggles to answer questions requiring technical knowledge or spatio-temporal reasoning.

Keywords

video question answering, human visual attention, multimodal large language model

1. Introduction

This paper presents a new VQA dataset based on demonstrating basic sewing machine operations. To our knowledge, THAVQA[1], which is also annotated with human visual attention, is the first task-oriented VQA dataset in German language.

The dataset building is a first step in the larger project aimed at developing an AI-assistant for a sewing machine workshop held at the Technische Hochschule Augsburg. This AI-assistant would support students when using sewing machines for the first time. For example, this could mean answering questions about basic machine settings or explaining fundamental sewing skills. Our dataset poses unique challenges for VQA models and is almost unique in the state-of-the-art VQA datasets since it is user- and task-oriented: the questions collected are those that a real user would ask for help while using the sewing machine. The process of operating the sewing machine was decomposed in a script into steps and sub-steps that were recorded and on which questions and answers were annotated. Specialized knowledge of the process and understanding of spatial and temporal relationships is required for answering the questions collected. In addition, the limited visual variety of the video scenes and the specialized language and dictionary challenge the models for VQA.

Annotating human attention in the video inputs of VQA models has recently been shown to improve their performance in user- and task-oriented datasets [1, 2]. In our dataset, the workshop instructor's eye gaze has been used as a proxy for human visual attention. The concept behind it is that visual human attention integrated as input into models for VQA can help the model distinguish between video frames, especially in datasets in which recorded scenes are very similar to each other as there are few participants and staged events.

Our paper also provides a first assessment on the VQA performance of the pre-trained MLLM Gemini 1.5 Pro[2] on THAVQA. Indeed, new releases of LLMs, such as Gemini 1.5 [3] but also GPT-4 [4], Llama 2 [5] or Claude 3 [6], now allow for visual inputs, making it possible to perform VQA tasks using pre-trained models directly.

To sum up, this paper presents (1) A new dataset with third-person videos of an instructor operating a sewing machine and first-person videos annotated with visual human attention, QA pairs in German, a script in German of the steps required to operate the machine; and (2) An evaluation of the performance of a pre-trained MLLM on generating open-ended answers from questions and videos of our dataset.

2. Related Work

The majority of state-of-art VQA datasets portray complex scenes composed of many events and participants, gathered using either synthetic simulation data or data sourced from movies, social media, video games or the web [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. VQA models are then tasked with answering questions about the

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

†These authors contributed equally.

moritz.kronberger1@tha.de (M. Kronberger);

viviana.ventura@tha.de (V. Ventura)

[1]https://github.com/tha-atlas/HowDoesSewingMachineWork.git

[2]https://deepmind.google/technologies/gemini/pro/

(a) (b) (c)

Figure 1: Video frames of the third-person (a) and first-person view with the human attention annotated as a circular outline (b) and an attention map (c).

videos' content. This requires a wide variety of reasoning abilities such as reasoning about spatial and temporal relationships, casual inference or relationships between actions and objects [16, 18].

In contrast, research on task-oriented VQA, where question answering supports users with tasks such as industrial assembly and disassembly [1, 2] or collaborative machine operation [19], is relatively limited. Similarly, the setting of our dataset, the tutorial on sewing machine operation, is task-oriented and requires specialized knowledge, which makes it difficult for pre-trained MLLMs to generate satisfactory answers from only their inherent knowledge. In line with the task-oriented approaches of Ilaslan et al. [1] and Gao et al. [10] we adopt both a fixed third-person view (TPV) and the first-person view (FPV) of the workshop instructor during the video recordings. To our knowledge no other German datasets exist specifically for task-oriented VQA.

Human and model attention in VQA seem to be related, as human visual attention has been shown to be correlated to model attention for VQA [20] and differences in their attention can be used to explain disagreement in VQA [21]. Human attention has been modeled explicitly by eye [1] and hand tracking [2] and included into the input of VQA models in order to highlight important parts of the videos that correspond to the user intentions. These annotations of human visual attention have been shown to improve VQA performance, even when using pre-trained encoders without specific fine-tuning to extract features from the visual data [1]. With these intuitions, we annotated the FPV videos in our dataset with human visual attention.

3. The Dataset

3.1. Dataset Structure

The setting of our custom VQA dataset is the introduction to sewing machine operation presented in a tutorial form. We based the contents on a sewing machine workshop held at the Technische Hochschule Augsburg as part of an elective module on Smart Textiles at the Faculty of Design. We first structured the contents and detailed instructions of the workshop in a script, which primarily served as a template for video data collection. The script contains seven larger tasks, such as setting up the machine and performing different kind of sewing operations on different kinds of fabrics, each with three to eight smaller sub-steps (35 in total), which in turn require multiple actions to be performed. The script's contents are available as part of the publicly accessible dataset (see Online Resources).

3.2. Video Data Collection

We recorded video data of the workshop being performed by the instructor. All videos depict a regular consumer-grade sewing machine being operated by the instructor at a table (see Figure 1). The video background is visually complex and reflects the real workshop environment. We also extended the video dataset to two student participants using exactly the same recording procedure (same environment, perspectives and script steps). The extended dataset, containing a total of 48 minutes of footage, is available on request. To reduce the chance of errors in the video demonstrations negatively impacting VQA performance, we rely exclusively on the expert demonstrations for the scope of this paper.

Two different camera perspectives were recorded simultaneously: a static TPV looking over the instructor's left shoulder towards the machine (see Figure 1a) as well as a dynamic FPV of the instructor (see Figure 1b). For recording the FPV we used the Tobii Pro Glasses 3 eye tracking glasses[3] and collected the instructor's eye gaze fixations for the entire duration of recordings. We split the recordings (TPV and FPV) into the 35 sub-steps and manually synchronized them across both perspectives.

We chose two different types of annotations to represent the human attention in FPV. First we annotated the 2D-location of the instructor's eye gaze via a red circular

[3] https://www.tobii.com/products/eye-trackers/wearables/tobii-pro-glasses-3

outline (FPV_C) (see Figure 1b), representing a bounding box for the current area of human attention, similar to the annotation style of Ilaslan et al. [1]. We also created a second annotation layer, attention maps (FPV_A), where each pixel is masked with increasing intensity with increasing distance to the gaze fixation point (see Figure 1c). Although this masking may obscure important information in the video, it clearly restricts the model's visual input to the human focal point.

3.3. QA Pair Collection

We recruited 10 German speaking crowdworkers on the Prolific[4] platform to formulate open-ended question-answer pairs on the recorded videos.[5] Crowdworkers were shown a random video in the TPV that represents a sub-step, together with the corresponding sub-step in the script. Giving annotators access to the script's contents, a description of the actions performed on the sewing machine by the instructor (see Section 3.1), did cause the resulting QA pairs to be less focused on the contents of the video and more focused on the contents of the textual descriptions. However, we still opted to include the textual context, in order to encourage the use of correct technical language by the non-expert annotators and to ensure a better understanding of the videos' contents. The resulting QA pairs were then manually annotated by reasoning type (see Figures 2-3 in the Appendix):

- *knowledge-based* reasoning when questions need technical knowledge to be answered;
- *spatial* reasoning when locations or directions are to be described;
- *temporal* reasoning when questions are related to the sequential order of actions;
- *perception-based* reasoning when the answer can only be retrieved by visually inspecting the video.

The categorization of QA pairs into these reasoning types is often ambiguous, especially when differentiating if a question pertains to knowledge-based reasoning as opposed to spatial or temporal reasoning. In fact most knowledge about how to sew is based on spatial and temporal information. For example the question of "What happens after winding the bobbin?" is temporal in nature but could also be answered from the model's inherent pre-training knowledge instead of extracting temporal information from the video input. We therefore approached the labeling process of QA pairs as follows:

- If a question can be answered by locating objects in the visual input it is categorized as requiring spatial reasoning.
- If a question can be answered by observing and relating the video input over multiple frames it is categorized as requiring temporal reasoning.
- If a question cannot be reasonably answered from the video input but rather requires using pre-training knowledge it is categorized as requiring knowledge-based reasoning.

This approach still leaves some amount of ambiguity, for example specialized knowledge about sewing-machine-specific terms may be required in order to identify the object, for example "the bobbin", to be located in a QA pair about temporal-based reasoning. For the QA pair annotation it was therefore decided if a question corresponds to a single reasoning type or if it should be assigned to multiple reasoning types.

The different reasoning types also give an indication of which dataset modality is required for the model to answer the dataset's questions. Strictly knowledge-based questions for instance primarily test the model's pre-training knowledge and are therefore not expected to profit from a visual input modality. Spatial and temporal questions both require the model to extract additional information from visual inputs. For spatial reasoning, a sequence of video frames might help with occlusion or depth perception, however, in most cases a static image will offer the required context for a spatial question to be answered. Temporal reasoning requires the model to relate visual information over a span of multiple frames, making video context a requirement to answer temporal questions.

Additionally, we discarded QA pairs that were either factually incorrect, not intelligible or ungrammatical.

3.4. Descriptive Statistics

In total the video recordings span 16 minutes and 24 seconds across the TPV and FPV with a mean duration of 14 seconds for single sub-step-related video clips.

Since the dataset's scenario only involves sewing machine operation, we expect limited variability within the contents of the videos. This might mean that the video data offers little usable information to a pre-trained MLLM. We quantified this lack of visual variation as the semantic similarity of video frames within a single video clip related to one of the 35 sub-steps. We obtained the semantic similarity scores by randomly sampling 20 frames for each clip and transforming them into embeddings using the CLIP model [22]. We used cosine similarity [23] as the distance metric and calculated the mean of the similarity matrix between all 20 embeddings. We compared this semantic similarity for the TPV and FPV, including both types of annotations for human visual attention (see Table 1). As expected, the frames within video clips are very similar, with the static TPV exhibiting the largest

[4] https://www.prolific.com

[5] Crowdworkers were offered an approximate hourly reward of 11.80€ including bonuses.

Table 1
Comparison of the mean semantic (cosine) similarity $[0, 1]$ of video frames within clips related to single sub-steps.

Perspective	Mean Semantic Similarity
TPV	0.97 ±0.01
FPV	0.93 ±0.02
FPV_C	0.93 ±0.02
FPV_A	0.94 ±0.02

Table 2
Mean statistics over single questions and answers as well as across all questions, answers and the entire dataset.

	Tokens	Lemmas	RTTR
Single questions	9.79 ±3.0	9.12 ±2.43	2.88 ±0.45
Single answers	12.58 ±8.74	10.45 ±5.83	2.99 ±0.85
Questions	1519	286	9.34
Answers	1950	371	9.94
Total	3469	502	10.31

semantic similarity between video frames. The FPV annotated with attention maps displays the second highest similarity score, possibly due to the fact that large portions of the frames are masked and the position of the focal point is not altering the embedding vector significantly. We do not find a difference between the similarity scores of the regular FPV and the FPV including the circle annotation of the eye gaze. Overall, this indicates that a pre-trained MLLM may struggle to extract and meaningfully interpret human attention information.

After manually filtering incorrect or unintelligible QA pairs and annotating the reasoning types we obtained a total of 122 QA pairs, with 1 to 9 QA pairs per sub-step of the script. Additionally, we prompted Gemini 1.5 Pro to answer the 122 questions, obtaining a total amount of 2562 answers, further details are described in Section 4. We found 96 QA pairs to pertain to knowledge-based reasoning, with 33 QA pairs requiring spatial-, 15 temporal- and 4 perception-based reasoning (see Figure 3 in the Appendix). A total of 24 QA pairs were annotated with more than one reasoning type due to ambiguity. All but one of these pairs was assigned the label "knowledge-based reasoning" in combination with at least one more reasoning type.

Additionally, we analyzed the diversity of QA pairs in terms of token and lemma counts as well as Root Type-Token Ratio (RTTR) calculated using the default parameters of Shen [24] (see Table 2). We calculated the descriptive statistics as a mean over singular questions and answers as well as across all questions, answers and the entire dataset. The questions and answers provided by the human annotators are largely brief and concise, resulting in low token and lemma counts alongside a low RTTR. When extending the calculations to all questions and answers or the entire dataset, repetitions become more frequent, evidenced by a higher RTTR.

4. Methodology

For the evaluation we selected Gemini 1.5 Pro[6] as an example of pretrained MLLMs. Gemini 1.5 Pro is part of a new family of highly-capable multi-modal models, Gemini 1.5, and it is a sparse mixture-of-expert Transformer-based model. Due to its is long input context of up to 10 million tokens it is capable of processing video inputs at a high resolution and sampling rate [3], giving it a good chance at extracting detailed visual information. We accessed Gemini through the Vertex AI inference API[7]. We prompted Gemini to answer the questions formulated by human annotators. To evaluate the model's performance, the answers generated by Gemini are manually compared against the human gold-standard answers. Two human annotators gave binary labels of whether or not the model answer could serve as an acceptable replacement for the human answer. The two annotators were trained by tagging part of the dataset together. Given the clarity of the binary annotation task, they proceeded to annotate the remaining part of the dataset by themselves. Instances where the model refused to answer due to a lack of information were labeled as not acceptable. For the final evaluation score we expressed the ratio of acceptable answers to the number of total answers as binary accuracy (see Table 3).

To evaluate the impact of different inputs (FPV, TPV, human visual attention, script) on the VQA performance of Gemini we constructed seven ablation settings:

First, we prompted the model with the questions and did not include any other context in form of textual information or videos. We refer to this ablation setting as the *naive baseline*. We expected this configuration to serve as the bottom limit of model performance, relying exclusively on the model's inherent knowledge gathered from pre-training.

For the second ablation scenario, we included the instructions for the sub-step of the script any given question was formulated for. These instructions do not only aid with knowledge-based questions but also contain important descriptions about the temporal order and spatial location of actions. Excluding perception-based reasoning, we therefore expected this ablation setting to represent the upper limit of model performance. As such, this ablation setting is referred to as the *text-only reference model*.

[6] https://deepmind.google/technologies/gemini/pro/
[7] https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/inference

Table 3
Mean binary accuracy of Gemini answers per ablation setting and reasoning type.

Ablation	Knowledge	Spatial	Temporal	Perception	All reasoning types
Naive baseline	0.36	0.61	0.29	0.25	0.42
TPV	0.42	0.61	0.38	**0.75**	0.48
FPV	0.43	0.71	0.27	0.67	0.5
FPV_C	0.4	0.71	0.27	0.58	0.48
FPV_A	0.43	0.68	0.29	0.5	0.49
Text-only reference model	**0.89**	0.76	0.56	0.08	0.79
Multimodal reference model	0.87	**0.84**	**0.71**	**0.75**	**0.84**

Third, we included a FPV video clip corresponding to the given question along with the sub-step instructions. We refer to this model as the *multimodal reference model* and expect it to perform similarly to the text-only reference model with the additional ability to reason about perception-based questions. If satisfactory answers cannot be generated from the model's pre-training knowledge, we would expect both reference models to outperform the naive baseline significantly.

In the remaining four ablation settings, we included a single video clip related to the given question with every prompt. Each ablation setting used video clips, either from a specific perspective (*TPV* or *FPV*) or a specific type of visual attention information, either the red circle (FPV_C) or the attention map (FPV_A). For these settings we did not include any other textual information, meaning all information present in the answers must have been inherent to the model or extracted from the video.

We repeated the same prompt for every question in every ablation setting three times to account for variations in the model's output. This resulted in 366 model responses per ablation setting, a total of 2562 answers. Additional information about the model prompts is provided in Section E of the Appendix. Since THAVQA is imbalanced towards knowledge-based questions, we reduced their amount by randomly sampling knowledge-based questions. We chose the sample size with a margin of error of 5%, a confidence of 95% and estimated the proportion maximally at 0.5. With finite population correction we therefore reduced the amount of knowledge-based model answers from 210 to 143 per ablation setting. Model answers including spatial reasoning accounted for 99, temporal reasoning for 45 and perception-based reasoning for 12 model answers per ablation setting. This means that the evaluated model answers were still imbalanced towards knowledge-based reasoning.

5. Evaluation

We calculated the binary answer accuracy (see Section 4) for every ablation setting and reasoning type as shown in Table 3. To test for statistical significance we calculated χ^2 in a contingency table of the binary "acceptable"-labels between every pair of ablation settings for every reasoning type. We accepted p-values < 0.05 as statistically significant.

Both reference models outperformed the naive baseline significantly in terms of total accuracy over all reasoning types ($4.28e^{-25} \leq p \leq 4.57e^{-19}$). This confirms that the chosen task-oriented VQA scenario of sewing machine operation was specialized enough, such that Gemini was not able to provide satisfactory answers using only its pre-training knowledge. For perception-based reasoning questions, no significant difference in accuracy between the naive baseline and the text-only reference model was found. However, both were outperformed significantly by the multi-modal reference model ($0.004 \leq p \leq 0.04$). We can therefore conclude that the model was generally able to extract meaningful information from the video inputs. Across all individual reasoning types other than perception-based questions, no statistically significant differences between the performances of the text-only and multi-modal reference model could be observed, indicating that the textual instructions included enough spatial and temporal information to make the additional video input redundant.

All video-only ablation scenarios (TPV, FPV, FPV_C, FPV_A) across all individual reasoning types except for perception-based reasoning were outperformed by both reference models, and did not show significant advantages over the naive baseline. Given that even the multimodal reference model was not able to significantly improve upon the text-only reference model, these results were to be expected. Similarly, the video-only ablation scenarios were able to improve over the accuracy of the naive baseline and the text-only reference model with respect to perception-based reasoning, although these results were above or close to the cutoff for statistical significance ($0.004 \leq p \leq 0.4$).

More importantly however, for any individual reasoning type, annotating human attention via both annotation types (FPV_C and FPV_A) did not significantly improve accuracy in comparison to the regular FPV or TPV videos. This confirms that the pre-trained MLLM was in fact

not able to meaningfully interpret the human attention annotations without fine-tuning.

Overall, the experimental setup was suitable to reveal differences in VQA performance for the different forms of video inputs and reasoning types. In fact, the task-oriented nature of THAVQA was challenging for a pre-trained MLLM such as Gemini: while the model was often able to extract enough information for questions requiring basic perception, this was not the case for questions involving complex reasoning about temporal or spatial dimensions that are peculiar of a procedural task such as sewing. For these types of reasoning the model achieved its best performances when detailed textual information related to the corresponding sub-steps was included in the ablation scenarios. Besides the nature of the questions formulated, maybe the videos are also challenging for the model: we can hypothesize that this is due to the high semantic similarity between the video frames, as we showed in Section 3.4.

5.1. Qualitative Analysis

If no video inputs were included for perception-based questions, such as retrieving the fabric's color, Gemini mostly pointed out that it was lacking the information required to provide an answer. Additionally, including video inputs seemed to help the model disambiguate questions. For example, the naive baseline misunderstood a question about removing excess threads from the work piece, interpreting it as referring to undoing entire unwanted seams. With video inputs, the model was able to infer that the question was simply related to trimming long threads hanging off the fabric. Finally, we found that video context seemed to encourage the model to provide descriptions of spatial relationships, even when this is not strictly required to answer the question.

Overall, we observed a positive effect of video inputs on the model's answers when compared to the naive baseline. Examples are provided in the Appendix (Figures 5- 7).

6. Conclusion

We provide a new task-oriented, German-language VQA dataset on demonstrations of sewing machine operation with open-ended human QA pairs and human visual attention: THAVQA. We then compared the VQA performance of Gemini 1.5 Pro on THAVQA varying the model inputs. We found that the task-oriented scenario of THAVQA was specific enough, such that the model could not rely on only its inherent knowledge to generate satisfactory responses. The questions contained in our dataset were over the capacity of the model to reason about the video data. Combining textual instructions with a first person video resulted in the best performing model across all reasoning types of questions.

When looking towards the design of a VQA model for a future, practical sewing machine assistant, video inputs could therefore be used mainly to improve the model's perception abilities, while a retrieval system for textual information could provide the necessary specialized knowledge.

Acknowledgments

This research was funded by the Bavarian State Ministry for Science and the Arts (StMWK: Bayerische Staatsministerium für Wissenschaft und Kunst - StMWK) as part of the Project "CHIASM" (Changenreiche industrielle Anwendungen für vortrainierte Sprachmodelle) and as part the High Tech Agenda of the Free State of Bavaria.

We thank Rebecca Bilger of the Education and Learning Lab for Sustainability Innovations (ELLSI) for her support with the topic of sewing machine operation, the scheduling and organization of data collection and her participation in the video dataset. We also thank the research group for Applied Technologies of Language and Assistance Systems (THA_atlas) at the Technische Hochschule Augsburg for supporting the project with advice and equipment.

References

[1] M. Ilaslan, C. Song, J. Chen, D. Gao, W. Lei, Q. Xu, J. Lim, M. Shou, GazeVQA: A Video Question Answering Dataset for Multiview Eye-Gaze Task-Oriented Collaborations, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 10462–10479. URL: https://aclanthology.org/2023.emnlp-main.648. doi:10.18653/v1/2023.emnlp-main.648.

[2] H. L. Tan, M. C. Leong, Q. Xu, L. Li, F. Fang, Y. Cheng, N. Gauthier, Y. Sun, J. H. Lim, Task-Oriented Multi-Modal Question Answering For Collaborative Applications, in: 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 1426–1430. URL: https://ieeexplore.ieee.org/document/9190659. doi:10.1109/ICIP40778.2020.9190659, iSSN: 2381-8549.

[3] Gemini Team, Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL: http://arxiv.org/abs/2403.05530. doi:10.48550/arXiv.2403.05530.

[4] OpenAI, GPT-4 Technical Report, 2024. URL: http://arxiv.org/abs/2303.08774. doi:10.48550/arXiv.2303.08774.

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. URL: http://arxiv.org/abs/2307.09288. doi:10.48550/arXiv.2307.09288, arXiv:2307.09288 [cs].

[6] Anthropic PBC, Introducing the next generation of Claude, 2024. URL: https://www.anthropic.com/news/claude-3-family.

[7] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, P. Chen, Y. Li, S. Lin, S. Zhao, K. Li, T. Xu, X. Zheng, E. Chen, R. Ji, X. Sun, Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis, 2024. URL: http://arxiv.org/abs/2405.21075. doi:10.48550/arXiv.2405.21075, arXiv:2405.21075 [cs].

[8] Y. Li, X. Chen, B. Hu, L. Wang, H. Shi, M. Zhang, VideoVista: A Versatile Benchmark for Video Understanding and Reasoning, 2024. URL: http://arxiv.org/abs/2406.11303. doi:10.48550/arXiv.2406.11303, arXiv:2406.11303 [cs].

[9] R. Rawal, K. Saifullah, R. Basri, D. Jacobs, G. Somepalli, T. Goldstein, CinePile: A Long Video Question Answering Dataset and Benchmark, 2024. URL: http://arxiv.org/abs/2405.08813. doi:10.48550/arXiv.2405.08813, arXiv:2405.08813 [cs].

[10] D. Gao, R. Wang, Z. Bai, X. Chen, Env-QA: A Video Question Answering Benchmark for Comprehensive Understanding of Dynamic Environments, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1655–1665. URL: https://ieeexplore.ieee.org/document/9711383. doi:10.1109/ICCV48922.2021.00170, iSSN: 2380-7504.

[11] A. Yang, A. Miech, J. Sivic, I. Laptev, C. Schmid, Just Ask: Learning to Answer Questions from Millions of Narrated Videos, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1666–1677. URL: https://ieeexplore.ieee.org/document/9710833. doi:10.1109/ICCV48922.2021.00171, iSSN: 2380-7504.

[12] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2630–2640. URL: https://ieeexplore.ieee.org/document/9009806. doi:10.1109/ICCV.2019.00272, iSSN: 2380-7504.

[13] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, D. Tao, ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 9127–9134. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4946. doi:10.1609/aaai.v33i01.33019127, number: 01.

[14] J. Lei, L. Yu, M. Bansal, T. Berg, TVQA: Localized, Compositional Video Question Answering, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1369–1379. URL: https://aclanthology.org/D18-1167. doi:10.18653/v1/D18-1167.

[15] Y. Jang, Y. Song, Y. Yu, Y. Kim, G. Kim, TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1359–1367. URL: https://ieeexplore.ieee.org/document/8099632. doi:10.1109/CVPR.2017.149, iSSN: 1063-6919.

[16] K. Yi*, C. Gan*, Y. Li, P. Kohli, J. Wu, A. Torralba, J. B. Tenenbaum, CLEVRER: Collision Events for Video Representation and Reasoning, 2019. URL: https://openreview.net/forum?id=HkxYzANYDB.

[17] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, S. Fidler, MovieQA: Understanding Stories in Movies through Question-Answering, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4631–4640. URL: https://ieeexplore.ieee.org/document/7780870. doi:10.1109/CVPR.2016.501, iSSN: 1063-6919.

[18] M. Grunde-McLaughlin, R. Krishna, M. Agrawala, AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 11282–11292. URL: https://ieeexplore.ieee.org/document/9577594. doi:10.1109/CVPR46437.2021.01113, iSSN: 2575-7075.

[19] T. Wang, J. Li, Z. Kong, X. Liu, H. Snoussi, H. Lv, Digital twin improved via visual question answering for vision-language interactive mode in human–machine collaboration, Journal of Manufacturing Systems 58 (2021) 261–269. URL: https://www.sciencedirect.

com/science/article/pii/S0278612520301217. doi:10.1016/j.jmsy.2020.07.011.

[20] E. Sood, F. Kögel, F. Strohm, P. Dhar, A. Bulling, VQA-MHUG: A Gaze Dataset to Study Multimodal Neural Attention in Visual Question Answering, in: A. Bisazza, O. Abend (Eds.), Proceedings of the 25th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2021, pp. 27–43. URL: https://aclanthology.org/2021.conll-1.3. doi:10.18653/v1/2021.conll-1.3.

[21] S. Hindennach, L. Shi, A. Bulling, Explaining Disagreement in Visual Question Answering Using Eye Tracking, in: Proceedings of the 2024 Symposium on Eye Tracking Research and Applications, ETRA '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1–7. URL: https://doi.org/10.1145/3649902.3656356. doi:10.1145/3649902.3656356.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html, iSSN: 2640-3498.

[23] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, USA, 2008.

[24] L. Shen, LexicalRichness: A small module to compute textual lexical richness, 2022. URL: https://github.com/LSYS/lexicalrichness. doi:10.5281/zenodo.6607007.

[25] M. Post, A Call for Clarity in Reporting BLEU Scores, in: O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, K. Verspoor (Eds.), Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 186–191. URL: https://aclanthology.org/W18-6319. doi:10.18653/v1/W18-6319.

[26] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

A. Online Resources

The dataset, including synchronized video data with annotated eye gaze as well as human formulated and model generated question-answer pairs with reasoning type annotations, is available via https://github.com/tha-atlas/HowDoesSewingMachineWork.git.

B. Crowdsourced Question-Answer Formulation

Figure 2: The question-answer formulation task as presented to human annotators.

C. Reasoning Types

Why do you use a zigzag stitch on elastic fabrics? Warum verwendet man einen Zickzack-Stich bei Elastischen Stoffen?	(a)
Where is the sewing machine's built-in thread cutter located? Wo befindet sich der integrierte Fadenschneider der Maschine?	(b)
What does the seamstress check at the end of the sewing? Was kontrolliert die Näherin am Ende des Nähens?	(c)
What color is the fabric in the video? Welche Farbe hat der Stoff in dem Video?	(d)

Figure 3: Questions requiring knowledge-based (a), spatial (b), temporal (c) and perception-based (d) reasoning.

D. Semantic Similarity of Human and Model Answers

We also evaluated the similarity between human and model answers for every ablation scenario as a sentence BLEU-score [25] and BERT-scores [26] with precision, recall and F1-score (see Table 4). However, we excluded these metrics from the main evaluation, since they do not provide a direct measure for the factual correctness of the model's responses. As expected, the reference model with access to the same textual information that annotators were using to formulate QA pairs achieves the highest semantic similarity to human answers.

E. Model Prompts

When including video data in the prompts, we found that Gemini had to be explicitly instructed to retrieve information from the video. We therefore also included information about the types of annotations for human visual attention in the prompt, where applicable, in order to increase the model's chances at recognizing the annotations. Additionally, we added a single few-shot example of the expected answer format in the prompt, without disclosing any factual information. We input the videos at full resolution. According to the Vertex AI documentation, videos in the prompts are sampled at one frame per second, with automated changes to the sampling rate being made in order to improve inference quality[8].

[8] https://ai.google.dev/gemini-api/docs/prompting_with_media#prompting-with-videos

Table 4
Mean BLEU- and BERT-scores (precision, recall and F1-score) between the human gold-standard and the model answers for each ablation scenario across all reasoning types.

Ablation	BLEU	Precision	Accuracy	F1
Naive baseline	4.39% ±4.11%	0.75	0.70	0.72
TPV	6.66% ±9.39%	0.75	0.72	0.74
FPV	6.23% ±7.52%	0.75	0.72	0.73
FPV_C	6.45% ±7.99%	0.75	0.72	0.73
FPV_A	6.34% ±8.01%	0.75	0.72	0.74
Text-only reference model	13.82% ±16.93%	**0.81**	0.76	0.78
Multimodal reference model	**16.74 ±21.5**	**0.81**	**0.78**	**0.80**

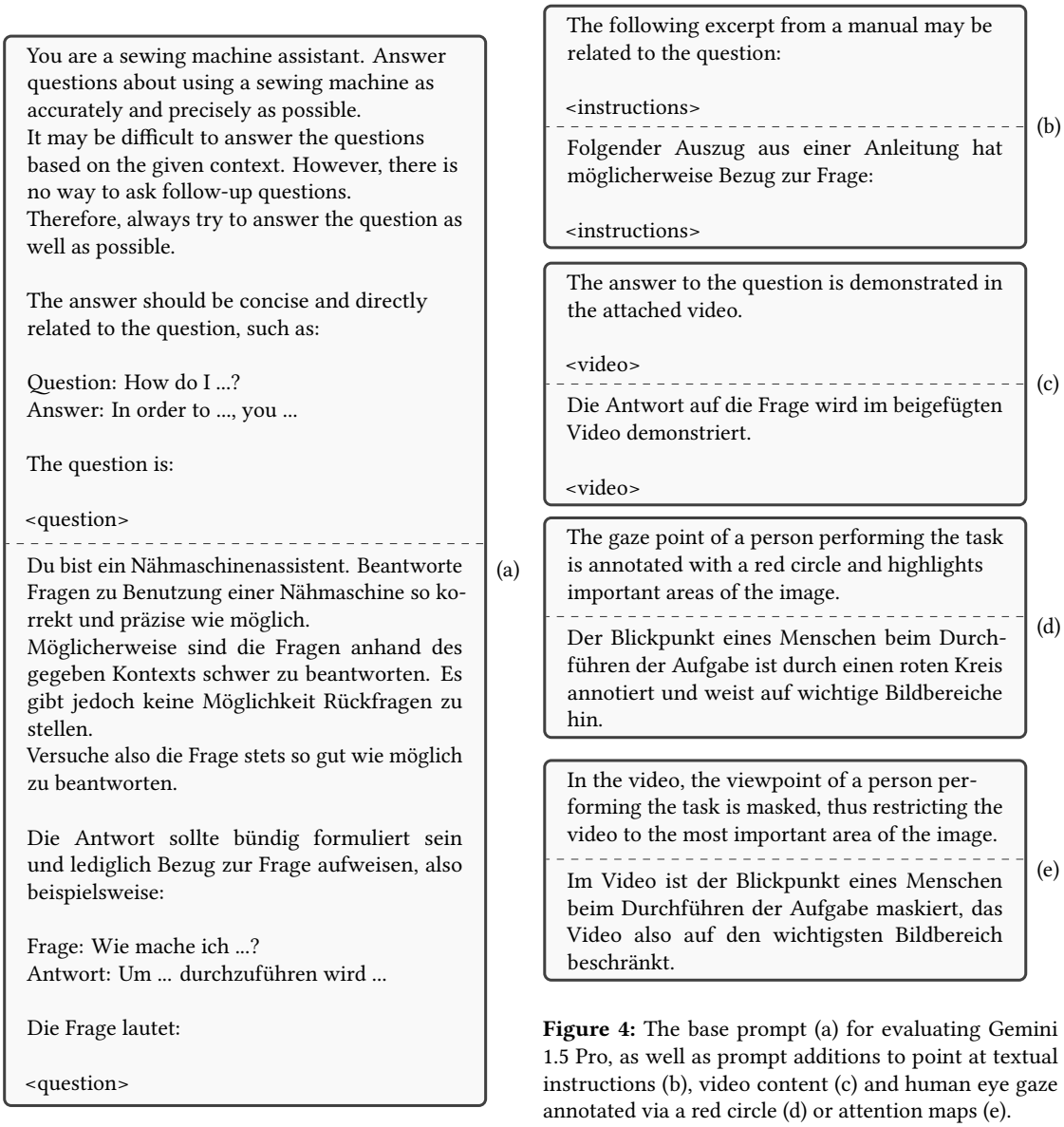

Figure 4: The base prompt (a) for evaluating Gemini 1.5 Pro, as well as prompt additions to point at textual instructions (b), video content (c) and human eye gaze annotated via a red circle (d) or attention maps (e).

F. Qualitative Analysis

Question: What color is the fabric in the video?

Welche Farbe hat der Stoff in dem Video?

Answer (naive baseline): It is impossible to determine the color of the fabric in a video without seeing the video.

Es ist unmöglich die Farbe des Stoffes in einem Video zu bestimmen, ohne das Video zu sehen.

Answer (FPV): The color of the fabric is blue.

Die Farbe des Stoffes ist blau.

Figure 5: Gemini recognizing lacking information for perception-based questions.

Question: How should excess threads be removed from the work piece?

Wie sollten die überschüssigen Fäden des Werkstückes entfernt werden?

Answer (naive baseline): To remove excess threads, use a seam ripper to carefully open the seams and remove the unwanted threads.

Um überschüssige Fäden zu entfernen, verwende einen Nahtauftrenner, um die Nähte vorsichtig zu öffnen und die unerwünschten Fäden zu entfernen.

Answer (FPV): To remove excess threads, cut them off with fabric scissors.

Um überschüssige Fäden zu entfernen, schneidet man diese mit einer Stoffschere ab.

Figure 6: Video inputs allowing Gemini to correctly interpret ambiguous questions.

Question: How do I manually insert the needle into the fabric?

Wie führe ich die Nadel manuell in den Stoff ein?

Answer (naive baseline): The needle can be manually inserted using the handwheel.

Mithilfe des Handrads kann die Nadel manuell eingeführt werden.

Answer (TPV): To manually insert the needle into the fabric, turn the handwheel *on the right side* of the sewing machine *towards yourself*.

Um die Nadel manuell in den Stoff einzuführen, dreht man das Handrad *an der rechten Seite* der Nähmaschine *zu sich heran*.

Figure 7: Gemini providing un-prompted spatial descriptions when video context is available.