Generalizable Detection of Student Engagement Markers in Online Learning Environments

Paper #2342

Abstract. Online learning has been on the rise since the COVID-19 pandemic due to its many advantages, however, the assessment of student engagement becomes much more difficult in this environment. Student engagement detection is crucial as it can be used by the teachers to alter their delivery (e.g., repeat, explain in a different way, change delivery speed, etc.). The emerging vision-language models (VLMs) are capable of detecting many different activities without extensive task-specific training. In this paper, we explore a method that finetunes a pretrained VLM to recognize student engagement markers. Our model learns to avoid incorrect answers during finetuning by using Direct Preference Optimization (DPO) on self-generated preference pairs. On publicly available student engagement datasets, our model shows superior performance over other approaches and substantially better generalizability over the traditional vision-based methods.

1 Introduction

With the development of network and multimedia technology, online learning environments have been available in education at all levels and its adoption has skyrocketed following the COVID-19 pandemic. Online learning is now well-entrenched and increasingly preferred because of its numerous benefits both to the students (e.g., avoidance of physical transportation) and the providers (lower cost and ability to handle large classes). However, online learning has some definite disadvantages due to the inability of the instructor to make face-to-face contact and assess comprehension or lack thereof. We use the word engagement to refer to the level of the students' focus and attention during the class. Research has demonstrated that student engagement provides a positive influence on academic achievement [25]. Keeping track of the student engagement level is beneficial for instructors since they can modify their lecture or instruction accordingly to keep the students' attention. In the online learning environment, automated monitoring of the student engagement level becomes essential because it is not possible for the instructor to visually scan the student video feeds manually and still be well engaged in the primary job of instruction.

In physical classrooms, student engagement can be assessed with some success based on physical tests and self-assessments that the education community has developed over time [24, 15]. In a remote education environment, we would like to measure the engagement non-intrusively and continuously using audio-visual means to provide discreet feedback to the instructors for real-time adjustments. Ideally, we would like to do the kind of assessment that experienced teachers can easily do in real classrooms by both observing the students and interacting with them verbally. For the purposes of this paper, we focus on the visual aspect, where the idea is to determine observable behavioral markers of each student and correlate it with their engagement level based on the context defined by the teacher's instructions or expectations. For example, a student looking intently at the screen would be considered engaged while the teacher is explaining something through slides, but it may be considered as a lack of engagement if the teacher's instruction is to perform a different activity, such as jotting down some notes. In this paper, we focus exclusively on accurately recognizing observable behavior.

Traditionally, visual behavior recognition has been done via specially designed deep learning models that must be trained on well-crafted labeled datasets. While such methods can do extremely well in recognizing the targeted behaviors, they are limited by the difficulties in model crafting and the creation of good quality and varied training datasets. They also tend to struggle in generalizing well to out-of-distribution samples, as shown in our experiments.

Large Vision-Language Models (VLMs) have recently been researched extensively to provide good descriptions (or "captions") of activities in images and short video segments. Numerous VLMs exist, both image-based (e.g., Clip [39], MiniGPT4 [52], LLAVA [31]) and video-segment based (e.g., X-Clip [32], Video-LLAMA [49], Video-chatGPT [33]). Like LLMs, VLMs can be specialized for specific applications via prompt engineering [31, 50] and finetuning [39, 11]. These models are claimed to be able to perform zero-shot prediction for a variety of visual recognition tasks by learning and mining the visual language correspondence between a large number of image-text pairs on the internet scale datasets such as CLIP [36] and ALIGN [19]. VLMs can potentially allow recognition of a wide range of visual behaviors with only a small number of finetuning examples. It is thus potentially much more generalizable in that a VLM system designed for detecting engagement markers for a specific environment can be updated for others such as: (1) recasting the system from classroom use to virtual business meetings, (2) updating the system for students in different age groups or grades (e.g., high school to middle school), (3) Specializing the system for use in a different country, etc.

Although a large number of observable features could be relevant for engagement, the goal of this paper is not to capture this breadth; instead, we take only a small set for which an adequate labeled dataset exists and focus on recognizing those as accurately as possible using a finetuned VLM and the analysis of the VLM output. We believe that our method can be easily adapted to accurately recognize a much larger set of visual behaviors.

In this paper, we explore MiniGPT-4 [52] and further finetune it for a few student engagement marker tasks. Additionally, to further enhance the generalizability of the model, we propose a novel Direct Preference Optimization (DPO) [37] method that finetunes the VLM using self-generated preference pairs. In particular, we design a specialized DPO finetuning method that yields a VLM capable

of accurately predicting student engagement markers from images (Sec. 3.2). Our design includes a self-generate pipeline to capture the broad distribution of the generated answers so that the preference pairs used to optimize the VLM can capture the actual distribution of incorrect answers. The proposed specialized DPO finetuning can be applied in any scenario where labeled dataset is available.

We compare our finetuning results with various ML models and show that it performs substantially better in predicting student engagement markers. We also evaluate the generalizability of our finetuned model to out-of-distribution samples by applying it to another unseen dataset which demonstrates good results.

The rest of the paper is organized as follows. Section 2 introduces the problem of student engagement and behavior markers indicating engagement levels and we also introduce some works that provide models to identify them. We also discuss relevant Vision-Language Models, advances in reinforcement learning approaches, and preference alignment optimization, as a basis to understand our approach. Section 3 describes the essential finetuning details of our method on the student engagement dataset. Section 4 explains the experimental setup, compares our method to other vision and VLM models, and analyzes the results. Finally, section 5 concludes the paper.

2 Background

2.1 Student Engagement

In this section, we briefly discuss various observable behavioral markers and their relevance to engagement.

An important behavioral aspect is the student's facial features and actions. Several works have attempted to characterize emotions from facial expressions and then translate them to engagement measures [17, 9, 43]. Although several deep learning vision models have enabled significant advances, these vision models often rely on expensive labeled data and train the network independently for each task. Thus, the student engagement detection task becomes a time-consuming and laborious process.

Student engagement and emotions are tightly correlated with learning [5, 13]. Emotions drive attention and attention drives learning [40]. Currently, there are only a few online learning studies that interactively account for the nature of learning, student motivation, and social/emotional learning [16, 8]. Establishing and maintaining student engagement is crucial, particularly as education shifts online. A significant challenge in online learning environments is the lack of support for students' emotions and engagement, where external distractions may cause disengagement. Facial expression datasets focused on facial expressions and head orientations often lack integration with online learning environments or the context created by teacher instructions. Notable facial expression datasets are the Affective-MIT Facial Expression [34] and Aff-Wild datasets [22, 21, 48]. Datasets that consider the student learning environment include the DAiSEE [17] and the Kaur [20] datasets.

There are many past studies but they define student engagement levels differently [10, 42]. Dewan et al. [10] proposed 3 ways of engagement detection, manual, semi-automatic, and automatic. The automatic method employs computer vision to analyze facial expressions for engagement using both part-based approaches that concentrate on specific facial areas and appearance-based approaches that consider the entire face. It also utilizes posture, gesture, and eye movement analysis, a widely recognized technique for assessing engagement. Sharma et al. [42] fuse information about eye and head movement with facial emotions to generate an engagement index that can be categorized as engaged, nominally engaged, and not engaged at all. They showed that there is a correlation between higher test scores and higher student concentrations. Researchers believe that machines can detect engagement levels just as humans are able to detect facial cues [43]. Other modalities can be used to detect student engagement. Altuwairqi et al. [3] used students' behavior, mouse movement, and keyboard keystrokes in addition to facial emotions to detect engagement. Huang et al. [18] also used features such as eye gaze direction, head pose, and eye coordinates, and achieved a higher accuracy than using these features individually. Abedi and Khan [1] framed the problem as a spatio-temporal classification problem on the DAiSEE dataset using Residual Network (ResNet) and Temporal Convolutional Network (TCN). With these prior works, a critical problem arises where the subject of attention that qualifies "engagement" is often not purely within the visual media itself. This makes the task of Student Engagement a subjective, qualitative task without a definitive context of reference.

2.2 Vision-Language Models

Vision-Language Models (VLMs) are multimodal models that fuse visual and textual information together to perform a task. There are two major avenues of research on VLMs. One avenue is understanding visual information from the visual and textual channels learned by VLMs and outputting textual descriptions. The other is outputting a relevant image from textual description. In this paper, we focus on the former.

There are commonly three modules of the VLMs: a vision module, a text module, and a fusion module that associates these two modalities. The vision module and text module are responsible for extracting essential features from the input image and text respectively. The fusion model then closely connects these features together through certain learning strategies. Techniques for building VLMs have evolved over time. Recent studies using transformer-based techniques have achieved state-of-the-art performance. Instead of using hand-designed feature descriptors or pre-trained word vectors/embeddings, transformer-based image and text encoders are employed in VLMs to learn image and text features individually or jointly.

Various techniques have been shown to get good performance and can learn the complicated connections between different modalities. CLIP [36], ALIGN [19], and DeCLIP [30] are vision-language models that bridge the visual and textual modalities by jointly training a text encoder and an image encoder using contrastive learning. These models minimize the loss between embeddings of image-text pairs that belong together and maximize the loss when they do not.

BLIP [26] designs a Multimodal mixture of Encoder-Decoder (MED). MED can either operate as an unimodal encoder, an image-grounded text encoder, or an image-grounded text decoder. One novelty of BLIP is that it combines three losses to train the model jointly. Besides the image-text contrastive loss (ITC), which is used to train the unimodal encoder in order to align the visual and textual features, BLIP uses image-text matching (ITM) loss to learn the difference between positive and negative image-text pairs. A language modeling (LM) loss is also used to generate captions for given images. To better associate visual and textual features, a cross-attention module is used to mix the visual features with the text features.

PrefixLM [38] is another approach to learning often used in pre-training. PrefixLM learns to predict the next word in the text sequence given the input part of the text as the prefix. In VLMs, it inputs an image and part of the text and learns to predict the text

sequence of words. It applies the same prefix concept to the image by utilizing a Vision Transformer (ViT) [12] that divides each image into patches and inputs the sequence of image patches in order into the model. The model obtains the visual embeddings by applying the convolution or linear projection over the image patches. Then, both the image embeddings and the text token embeddings which are converted from the text prefix are inputted into the transformer blocks. SimVLM [47] carries out such an architecture using the PrefixLM for Vision Language Pretraining (VLP).

In addition, learning image embeddings that align with a frozen language model has proved to be more effective in learning a new language task with only a few examples. That is, instead of learning both visual and textual modules, the model uses a pretrained language model without finetuning and only learns how to align the visual features with the text token embeddings by updating the parameters of the image encoder. Several VLMs employ such architecture, such as Frozen [46] and ClipCap [35].

Furthermore, Flamingo [2] uses pretrained vision encoder and the LLM for few-shot learning. Flamingo adds new cross-attention layers between the frozen language model layer to tune the language model layer based on the visual input. The approach taken in Flamingo allows it to interleave text, video, and images. Either an image or a video is passed into the image encoder and it uses the perceiver resampler module to convert it into tokens.

BLIP-2 [27] uses Q-Former, which is a lightweight transformer, to bridge the pretrained frozen image encoders and pretrained frozen LLMs. BLIP-2 uses two-stage pretraning for Q-Former. The first stage pretraining learns the vision-language representations. Q-Former sets the learnable embeddings query which extracts visual features (most relevant to the corresponding text) from the input image. Like BLIP, BLIP-2 sets three optimization objectives, image-text matching, image-grounded text generation, and image-text contrastive learning. To prevent information leaks, a different attention mask is employed for different objectives. The second stage pretraining learns to output visual representations that can be interpreted by the frozen LLM. A fully connected layer is used to convey the Q-Former output to the same dimension as the chosen LLM input.

MiniGPT-4 [52] is a VLM that builds on ViT, Q-Former, and open-source LLMs (e.g. Vicuna [7]). Both the visual and language components are frozen. The frozen components are bridged through a trainable linear layer. We discuss how the model is pretrained and finetuned in section 3. This model is selected as a base of our method since the model can achieve vision language abilities comparable to GPT-4 while being amendable to finetuning with low computational requirements.

2.3 Reinforcement Learning with Human Feedback

LLMs and VLMs are challenging to train as good output from these models is hard to define. Reinforcement Learning with Human Feedback (RLHF) provides a way for humans to be a source of the reward model without explicitly modeling the rewards. In LLMs and VLMs, RLHF is used to finetune the models, getting humans to choose between various outputs and learning a model to estimate the rewards. This is used to tune the LLMs and VLMs.

Pretaining stage trains the language model to predict the next token given the prior text information using the large collection of datasets for natural language processing tasks. The loss used in the pretraining stage is normally the cross entropy loss.

Supervised Fine-tuning (SFT) stage finetunes the pre-trained language model on datasets for specific downstream tasks. The datasets are normally high-quality datasets with appropriate instructions (prompts) and reasonable responses. The model obtained after the SFT finetuning is demoted as π_{ref}.

Preference sampling and Reward Learning: For a dataset D, for each input x, there is a pair of preferred answers (y_ω, y_ι), where y_ω denotes the answer that human labelers expressed preference for and y_ι denote the dispreferred answer. We can model this by a latent reward model r^* that generates the underlying preferences. A commonly used approach to model preferences is assuming the probability $p^*(y_\omega \succ y_\iota|x)$, which represents the probability of preferring answer y_ω over answer y_ι, as a sigmoid σ of the reward difference as shown in equation1.

$$p^*(y_\omega \succ y_\iota|x) = \sigma(r^*(x,y_\omega) - r^*(x,y_\iota)) \quad (1)$$

We need to use a reward model r_ϕ to estimate the true reward of human preference via maximum likelihood since the true reward function is not accessible. For this, we minimize the negative log-likelihood loss.

$$\mathcal{L}_R(r_\phi, D) = -\mathbb{E}_{x,y_\omega,y_\iota\sim D}[\log\sigma(r_\phi(x,y_\omega) - r_\phi(x,y_\iota))] \quad (2)$$

Reinforcement Learning Optimization: This phase uses the learned reward function to further optimize the language model. To prevent model collapse and maintain the proximity to the distribution of reference model π_{ref}, a KL divergence penalty is added.

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)}[r_\phi(x,y)] - \beta\mathbb{D}_{KL}[\pi_\theta(y|x) \,\|\, \pi_{ref}(y|x)]$$

where π_θ is the policy model that RLHF is optimizing, and β is the hyperparameter to restrict how far the model can deviate from the base reference model. Due to the non-differentiability, the objective is normally optimized with an RL algorithm such as Proximal Policy Optimization (PPO) [41]. PPO is a policy gradient algorithm that uses a novel objective function to find the best policy. The novel objective function minimizes the difference between the new and old policy so that PPO avoids too large a policy update which may destabilize the learning.

2.4 Multi-Modal Preference Alignment

There are a few works that study preference alignment optimization for Multi-Modal VQA tasks. SILKIE [28], HA-DPO [51], and Li et al. [29] focus on investigating the effects of closed form preference alignment methods and explore methodologies to build preferred and dispreferred pairs. HA-DPO and SILKIE both use GPT-4V while Li et al. uses Gemini to give preference annotation. HA-DPO also introduces an approach to utilize to formulate style-consistent positive and negative pairs.

3 Methodology

We propose a generalizable detection method that utilizes pretrained vision-language models to identify the student engagement markers. In this paper, we build our model based on MiniGPT-4 [52] which aligns visual information from a frozen vision encoder with a frozen advanced large language model (LLM) using one projection layer. We finetune MiniGPT-4 on our dataset using a set of prompts that are more relevant to the student engagement problem. In order to achieve more general performance, we optimize the MiniGPT-4 model with Direct Preference Optimization (DPO) [37] using self-generated preference pairs.

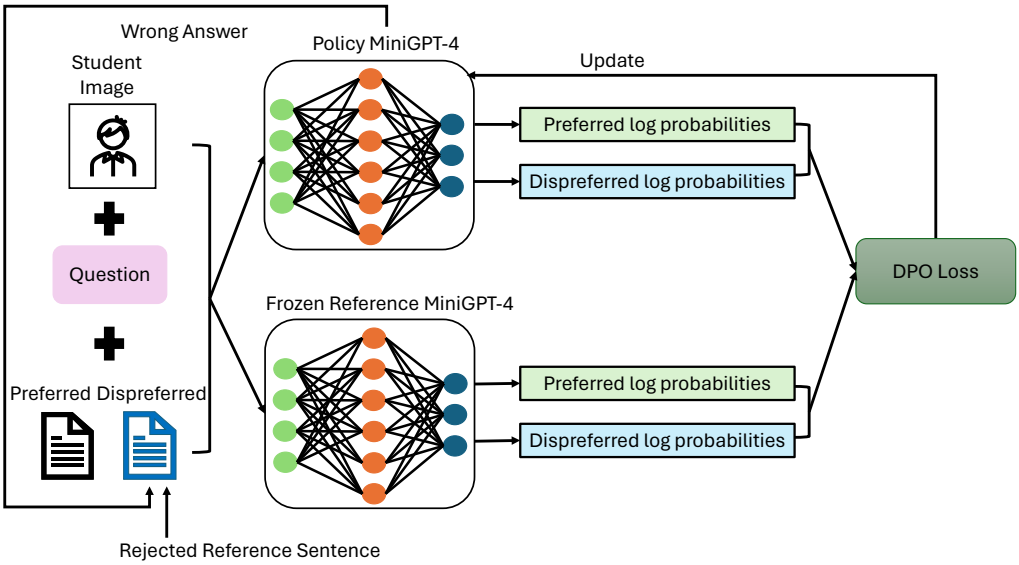

Figure 1. Architecture of DPO based Finetuning with self-generated preference pairs.

3.1 Finetuning MiniGPT-4 with Student Engagement Dataset

The vision encoder employed in MiniGPT-4 consists of a ViT [12] backbone and its pre-trained Q-Former [27]. For the language decoder, MiniGPT-4 utilizes the Vicuna [7] and Llama2 [45] model which is constructed upon LLaMA [44].

MiniGPT-4 adopted a two-stage training process. The **first stage** is the pretaining stage which aims to obtain the vision language knowledge. The dataset used in the pretaining stage is a combined large collection of the aligned image-text dataset of Conceptual Caption, SBU, and LAION. However, after this stage, the outputs that are generated may contain incoherent linguistic outputs, such as repetitive words or sentences, fragmented sentences, or irrelevant content.

The **second stage** alignment process utilizes the image-text pairs, in which the image descriptions are automatically generated by MiniGPT-4 and then mended by ChatGPT, to finetune the pretrained model of the first stage. The prompt used follows a similar structure to the one that is used in the Vicuna language model. This is used to encourage MiniGPT-4 to generate more detailed descriptions to form image-text pairs. The prompt is shown as follows:

###Human: <Img><ImageFeature></Img>Describe this image in detail. Give as many details as possible. Say everything you see. ###Assistant:

Then post-processing is performed to fix the noisy or incoherent descriptions that are possible in the MiniGPT-4 model after the first pretraining stage. The post-processing includes using ChatGPT to mend the descriptions and manually verify if the descriptions are correct for the corresponding images.

After the post-processing of image-text pairs, MiniGPT-4 is finetuned with the original prompts following the template below: *###Human: <Img><ImageFeature></Img> <Instruction>###Assistant:* The $< Instruction >$ is randomly sampled from a predefined instruction set. The original instruction set contains different forms of instructions as shown below:

- *Describe this image in detail.*
- *Take a look at this image and describe what you notice.*
- *Please provide a detailed description of the picture.*
- *Could you describe the contents of this image for me?*

After the second stage, The output descriptions generated by MiniGPT-4 of the corresponding images become more natural, reliable, and human-friendly compared to the first stage.

3.1.1 Student Engagement Detection Finetuning

In our model, we introduce a different finetuning that aims to generate specific image descriptions to describe behaviors that indicate different engagement levels of students. We use the Student Engagement Dataset (SED) [9] which has 3 different categories of student engagement markers: 'looking at their paper', 'looking at their screen', or 'wandering'. According to these different labels, we set the corresponding reference sentence as follows:

- *The person is looking down at the paper*
- *The person is looking straight at the screen*
- *The person is looking away*

Accordingly, we modify the instruction set of prompts from the general instruction to the questions that are used specifically to determine the students' behavior. The questions used in our task are shown below:

- *Is the person looking straight at the screen?*
- *Is the person looking down at the paper?*
- *Is the person looking away?*
- *Is the person looking straight at the screen? Is the person looking down at the paper? Is the person looking away?*

During the finetuning, one of these questions is used randomly. We found that using questions in this way produced a higher baseline. The reasons for this are perhaps worthy of further investigation. The loss we used for the finetuning is the same as the LLaMA model. We calculate the cross entropy loss between the tokens generated by our model and the labels that we targeted. We shift the generated tokens and the targeted labels so that the model predicts the next token based on the previous one.

We evaluated two different versions of MiniGPT-4 model in our experiments. One version is that we perform our finetuning using the MiniGPT-4 (Vicuna) checkpoint. Another version is that we perform our finetuning after using the MiniGPT-4 (Llama2) checkpoint. The MiniGPT-4 (Vicuna) achieves an accuracy of 95.2% after our finetuning whereas MiniGPT-4 (Llama2) yields 88.6%. Therefore, we use MiniGPT-4 (Vicuna) as our base model.

3.2 Direct Preference Optimization using Self-generate Preference Pairs

Direct Preference Optimization (DPO) provides a simplification of traditional feedback-aligned LLMs. DPO avoids training a Reinforcement Learning (RL) loop as RLHF. Instead, DPO proposes a closed-form loss that leverages a particular choice of reward model parameterization in equation 2. Equation 3 shows the detail of DPO loss, where y_ω represents the preferred answer, y_ι represents the dispreferred answer, π_θ represents the policy model we are optimizing, and π_{ref} represents the reference model that we constraint with. DPO loss updates mainly maximize the margin between preferred and dispreferred answer pairs as shown in equation 4.

$$L_{DPO}(\pi_\theta,\pi_{ref}) = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_\omega|x)}{\pi_{ref}(y_\omega|x)}-\beta\log\frac{\pi_\theta(y_\iota|x)}{\pi_{ref}(y_\iota|x)}\right)\right] \quad (3)$$

$$r_\omega = \log(\pi_\theta(y_\omega|x))-\log(\pi_{ref}(y_\omega|x))$$

$$r_\iota = \log(\pi_\theta(y_\iota|x))-\log(\pi_{ref}(y_\iota|x))$$

$$L_{DPO}(\pi_\theta,\pi_{ref}) = -\mathbb{E}[\log\sigma(\beta(r_\omega-r_\iota))] \quad (4)$$

The datasets that are used in academic research areas of feedback-aligned language models are preference data pairs such as HH [4], SHP [14], and OASST [23]. These datasets are collected from conversation questions/instructions and responses annotated with human preferences. However, for the student engagement detection problem, we normally work with multi-class labeled datasets. Various labels in the datasets indicate the different engagement behaviors of the student. Therefore, we need to transform the labeled datasets to preference dataset $D = \{x^i,y^i_\omega,y^i_\iota\}^N_{i=1}$, where N denotes the number of data samples, where for each input image x, we assign an answer pair of preferred answer and dispreferred answer (y_ω,y_ι). A simple approach is to utilize the given ground truth to generate a preference dataset by using the given label of each image as the preferred answer and using the other labels in the dataset as the dispreferred answer. However, this approach will introduce bias to the vision-language model since the distribution of possible dispreferred answers has a much wider range.

In order to yield better answers, it is reasonable to extend the reference data pairs D so that the dispreferred answer comes from a more general distribution than ground truth. Since we do not have the distribution of generated answers of the vision-language model, **we use the wrong answers generated from the vision-language model itself to extend the dispreferred answers**. In practice, we form the dispreferred answers during training in the following manner. The architecture of the DPO based finetuning is shown in Figure 1. For $a\%$ of the time, we check the generated answers from Policy MiniGPT-4 and use the wrong answers as the dispreferred answers. For the rest of the time, we randomly choose from other labels besides the correct label as the dispreferred answers.

4 Performance Evaluation

4.1 Experimental Setup

4.1.1 Datasets

Our evaluation utilizes two public datasets, the Student Engagement Dataset (SED) [9] and the DAiSEE dataset [17]. SED contains both an unbalanced and balanced dataset. The unbalanced dataset contains 18,721 frames which are sampled at one frame-per-second (FPS) from 400 videos collected from 19 students. The balanced dataset is a smaller version that removes similar samples for each of the three classes, resulting in a more balanced number of samples between 'looking at their paper', 'looking at their screen', or 'wandering'. The balanced dataset contains 1973 frames. Both the unbalanced dataset and balanced dataset contain the three categories of student data.

DAiSEE dataset is a large labeled student engagement level dataset that is collected by a web camera during the period of a student watching educational and recreational videos. The dataset contains 9068 video snippets collected from 32 female and 80 male subjects of age 18 to 30. There are 6 different location settings (dorm rooms, crowded lab spaces, library, etc) and 3 different illumination settings (light, dark, and neutral) of the video environment in the DAiSEE dataset. The dataset is originally labeled with 4 different student engagement levels: boredom, confusion, engagement, and frustration. In our evaluation, we sampled the videos at frame level and randomly selected 1046 frames out of them. Then we relabeled these frames with the three SED categories. The idea is to evaluate whether the model can apply the knowledge learned from the student engagement dataset to an out-of-distribution dataset under the same premise that both the DAiSEE and SED datasets have students in front of a web camera. With this premise, we annotated the DAiSEE dataset according to the evaluation framework of SED with labels 'looking at their paper', 'looking at their screen', and 'wandering'.

We have also identified 85 samples (hard samples) from SED that were selected based on misidentification by MiniGPT-4, which we call the handpicked SED dataset. These hard samples tend to contain images of students that face one direction whilst their gazes face another. We further evaluated GPT-4V (OpenAI (2023)) as opposed to our finetuning methodology on this handpicked dataset.

4.1.2 Preprocessing

We used 80% of the data samples from the balanced SED dataset for finetuning and 20% of the samples for our evaluation with balanced data. We also evaluated our model with the raw SED dataset. We excluded the training balanced sample from the raw SED dataset and used the rest of the samples for evaluation. For both training and testing data samples, we combined the labels for the various categories and created an annotation JSON file that is a collection of data pairs where each image ID is associated with a reference sentence based on the category of that image accordingly. The detail of the reference sentence is discussed in section 3.1. We also compared the results of our finetuning prompts with the original prompts used by MiniGPT-4. Additionally, the evaluation results, using the original MiniGPT-4 checkpoint without finetuning on the SED dataset, are shown as the baseline of the MiniGPT-4 model.

For the evaluation of the DAiSEE dataset, we manually labeled 1046 frames with the three SED categories, assigned image IDs for each frame, and constructed an annotation JSON file. The annotation file for this dataset was constructed in a similar fashion to SED.

	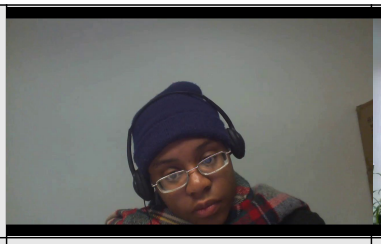	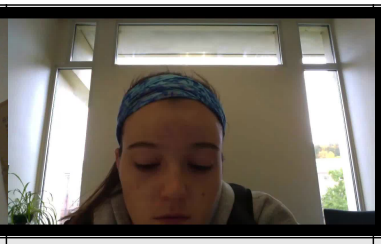	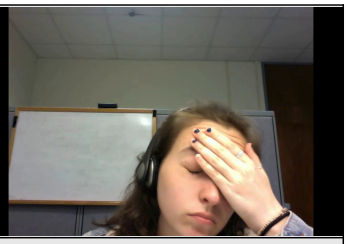
Reference sentence	The person is looking straight at the screen	The person is looking down at the paper	The person is looking away
MiniGPT-4 & DPO	The person is looking straight at the screen	The person is looking down at the paper	The person is looking away
MiniGPT-4 (SED specific prompts)	The person is looking straight at the screen	The person is looking down at the paper	The person is looking straight at the screen
MiniGPT-4 (Original prompt)	The person is looking straight at the screen	The person is looking straight at the screen	The person is looking away
MiniGPT-4 (Original model)	The person is wearing a blue beanie and a red scarf and has a pair of headphones on. The person is looking down at the paper on the desk in front of them. The image is a close up of the person's face, with the headphones on their ears and the paper in front of them.	The person is looking straight at the screen. They are not looking down at the paper or away from the screen.	1. The person is looking straight at the screen. 2. The person is looking down at the paper. 3. The person is looking away from the camera.

Figure 2. Comparison of the generated sentences using four different MiniGPT-4 based finetuning models for the three student engagement behavior markers. The image samples are selected from the SED dataset.

Method	Acc	F1	SS Acc	SS F1
MiniGPT-4 & DPO	96.7	96.7	96.7	96.7
MiniGPT-4 (SED specific prompts)	95.2	95.2	95.2	95.2
MiniGPT-4 (Original prompt)	87.6	87.4	87.6	87.4
MiniGPT-4 (Original model)	58.6	58.9	40.2	39.7
MobileNet (Pretrained ImageNet)	94	-	×	×
Xception (Pretrained ImageNet)	88	-	×	×
VGG16 (Pretraiend ImageNet)	85	-	×	×
Head Pose Estimator (Logistic Reg.)	60	-	×	×
Head Pose Estimator (Conditional)	55	-	×	×

Table 1. Overview of evaluation results (%) on balanced SED dataset.

Method	Acc	F1	SS Acc	SS F1
MiniGPT-4 & DPO	94.6	95.8	94.6	95.8
MiniGPT-4 (SED specific prompts)	90.8	92.8	90.8	92.8
MiniGPT-4 (Original prompt)	89.5	90.0	89.2	89.8
MiniGPT-4 (Original model)	50.5	61.2	48.6	57.8
MobileNet (ImageNet Pretrained)	89.9	91.3	×	×
Xception (ImageNet Pretrained)	87	87.9	×	×

Table 2. Overview of evaluation results (%) on raw SED dataset.

Method	Acc	F1	SS Acc	SS F1
MiniGPT-4 & DPO	88.4	87.9	88.5	88.0
MiniGPT-4 (SED specific prompts)	87.1	86.9	87.1	86.9
MiniGPT-4 (Original prompt)	88.0	87.7	87.2	87.2
MiniGPT-4 (Original model)	62.2	70.7	56.7	67.3
MobileNet (ImageNet Pretrained)	26.7	33.2	×	×
Xception (ImageNet Pretrained)	55.9	65.5	×	×

Table 3. Overview of evaluation results (%) on DAiSEE dataset.

Method	Acc	F1	SS Acc	SS F1
MiniGPT-4 & DPO	84.7	84.2	84.7	84.2
MiniGPT-4 (SED specific prompts)	81.2	80.8	81.2	80.8
MiniGPT-4 (Original prompt)	70.6	72.1	70.6	72.1
MiniGPT-4 (Original model)	56.5	61.3	31.8	34.7
GPT-4V	74.2	72.6	×	×
MobileNet (Imagenet Pretrained)	82.3	83.5	×	×
Xception (ImageNet Pretrained)	84.7	85	×	×

Table 4. Overview of evaluation results (%) on handpicked SED dataset.

4.2 Results

The following prompt is used when evaluating all the models:

Give the following image: <Img>ImageContent</Img>. You will be able to see the image once I provide it to you. Please answer my questions.###Human: <Img><ImageFeature><Img>Is the person looking straight at the screen? Is the person looking down at the paper? Is the person looking away?###Assistant:

We obtain the correctness of the generated results using two different methods, keyword evaluation and sentence similarity (SS) evaluation. For the keyword evaluation, we consider the output as the correct answer if the generated sentence contains the desired keywords (i.e. paper, screen, away) representing the reference sentence. For the sentence similarity evaluation, we use a pretrained sentence transformer to determine whether the generated sentence conveys the same meaning as the reference sentence. We do this by comparing the embedding of the generated sentences and the three candidate reference sentences using an SS transformer. If the reference sentence has the highest similarity score among all three possible reference sentences, it is marked as correct, otherwise, it is marked as incorrect.

The sentence transformer we used in the sentence similarity evaluation to compute the sentence embeddings is BGE-M3 [6]. We use cosine similarity to capture the semantic similarity.

The motivation for introducing sentence similarity is that the keyword evaluation is not able to properly capture the correctness of the generated results if the outputs contain the keywords in a negative sentence or the outputs contain multiple keywords. Figure 2 shows the generated results of different MiniGPT-4 based finetuning models. For example, the output result generated by MiniGPT4 (Original model), "The person is looking straight at the screen. They are not looking down at the paper or away from the screen.", should be counted as "The person is looking straight at the screen" instead of other categories. A keyword approach would incorrectly categorize the response.

We evaluate the performance of our model using both the accuracy and F1 score obtained by keyword and SS evaluation to capture correctness. We compare the results of our model (MiniGPT-4 & DPO) against the MiniGPT-4 model finetuned using SED prompts, MiniGPT-4 model finetuned using the original MiniGPT-4 prompts, and the original MiniGPT-4 model provided by the author. We also compare it with the original deep learning vision model results. We finetuned MobileNet and Xception on the balanced SED dataset and obtained similar accuracy results on the balanced SED dataset to those presented in [9]. We note that it makes no sense to compute the sentence similarity results for these models since they directly output one of the three classes. As stated previously, we finetuned our model on the balanced SED dataset, and evaluated it using the balanced SED dataset, the raw SED dataset, the DAiSEE dataset, and the handpicked SED dataset.

We show the evaluation results on the balanced SED dataset in Table 1. As we can see from this table, our method achieved 96.7% accuracy and F1 score, substantially better than the ones by the original deep learning vision models. The results for MobileNet, Xception, and VGG16, and the two head pose estimators are taken from [9]. Since they only report the accuracy, the other results for these methods are left blank. In all the tables, a $\times$ means that the sentence similarity result is not relevant to that specific model.

On the raw SED dataset, our model outperformed the other models across all four measures (see Table 2). On the unbalanced raw dataset, our model performs a bit worse compared to itself on the balanced dataset. Interestingly, the results of MiniGPT-4 (Original prompt) outperform itself on the unbalanced raw dataset compared to the balanced dataset. We suspect that even though MiniGPT-4 (Original prompt) is trained on the balanced dataset, it is still skewing toward the predominant class found in the raw unbalanced dataset.

Table 3 shows the evaluation results on DAiSEE dataset, and it shows an SS accuracy of 88.5% on this out-of-distribution dataset and it just out-competes the other finetuned MiniGPT-4 model on all four measures. In contrast, the results of MobileNet and Xception drop significantly, which clearly shows that they are not able to generalize well to out-of-distribution samples.

On the handpicked SED dataset which contains the hard samples, our model outperformed GPT-4V, and it shows an accuracy of 84.7% while GPT-4V shows an accuracy of 74.2%. We still maintain superior performance compared to other MiniGPT-4 based models. At the same time, we find it interesting that the hard samples cause all the MiniGPT-4 models to perform worse. This performance hit does not transfer to the purely vision models.

5 Conclusions and Future Work

In this paper, we focused on the task of accurate recognition of engagement relevant visual behavior markers using VLMs. We exploited the direct performance optimization (DPO) approach and proposed a modification to its finetuning that uses the model's responses to strengthen its performance. Unlike other preference alignment models, the proposed DPO finetuning generates preference data pairs using the wrong answers generated by the policy model during finetuning. This approach leverages pretrained VLMs and benefits from the large dataset it is trained on and finetunes to the student engagement domain. This makes the approach easily extensible for recognizing a large variety of visual markers relevant to engagement. We showed that our model's performance is superior to both purely vision models and other VLMs. We also demonstrated the generalizability with an out-of-distribution dataset, DAiSEE. In the future, we will apply it to other visual markers such as recognition of facial expressions, postures, body movements, etc. We will also study how these markers can be translated into engagement measures in a context-specific way.

References

[1] A. Abedi and S. S. Khan. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network, 2021.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] K. Altuwairqi, S. Kammoun jarraya, A. Allinjawi, and M. Hammami. Student behavior analysis to measure engagement levels in online learning environments. *Signal, Image and Video Processing*, 15, 10 2021. doi: 10.1007/s11760-021-01869-7.

[4] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[5] R. S. Baker, S. K. D'Mello, M. T. Rodrigo, and A. C. Graesser. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241, 2010. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2009.12.003. URL https://www.sciencedirect.com/science/article/pii/S1071581909001797.

[6] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[7] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[8] N. B. S. Danielle S. McNamara, Eileen Kintsch and W. Kintsch. Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14(1):1–43, 1996. doi: 10.1207/s1532690xci1401_1. URL https://doi.org/10.1207/s1532690xci1401_1.

[9] K. Delgado, J. M. Origgi, T. Hasanpoor, H. Yu, D. Allessio, I. Arroyo, W. Lee, M. Betke, B. Woolf, and S. A. Bargal. Student engagement dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3628–3636, October 2021.

[10] M. A. A. Dewan, M. Murshed, and F. Lin. Engagement detection in online learning: a review. *Smart Learning Environments. Smart Learning Environments*, 6(1):1–20, 2019. doi: 10.1186/s40561-018-0080-z.

[11] X. Dong, A. T. Luu, M. Lin, S. Yan, and H. Zhang. How should pre-trained language models be fine-tuned towards adversarial robustness? *Advances in Neural Information Processing Systems*, 34:4356–4369, 2021.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al.

An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] S. D'Mello, B. Lehman, R. Pekrun, and A. Graesser. Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170, 2014. ISSN 0959-4752. doi: https://doi.org/10.1016/j.learninstruc.2012.05.003. URL https://www.sciencedirect.com/science/article/pii/S0959475212000357.

[14] K. Ethayarajh, Y. Choi, and S. Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

[15] K. A. Fuller, N. S. Karunaratne, S. Naidu, B. Exintaris, J. L. Short, M. D. Wolcott, S. Singleton, and P. J. White. Development of a self-report instrument for measuring in-class student engagement reveals that pretending to engage is a significant unrecognized problem. *PloS one*, 13(10):e0205828, 2018.

[16] A. C. Graesser. Deeper learning with advances in discourse science and technology. *Policy Insights from the Behavioral and Brain Sciences*, 2 (1):42–50, 2015. doi: 10.1177/2372732215600888. URL https://doi.org/10.1177/2372732215600888.

[17] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian. Daisee: Towards user engagement recognition in the wild, 2022.

[18] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang. Fine-grained engagement recognition in online learning environment. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 338–341, 2019. doi: 10.1109/ICEIEC.2019.8784559.

[19] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[20] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall. Prediction and localization of student engagement in the wild, 2018.

[21] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou. Recognition of affect in the wild using deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1972–1979, 2017. doi: 10.1109/CVPRW.2017.247.

[22] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6–7):907–929, Feb. 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01158-4. URL http://dx.doi.org/10.1007/s11263-019-01158-4.

[23] A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z.-R. Tam, K. Stevens, A. Barhoum, N. M. Duc, O. Stanley, R. Nagyfi, S. ES, S. Suri, D. Glushkov, A. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, and A. Mattick. Openassistant conversations – democratizing large language model alignment, 2023.

[24] E. S. Lane and S. E. Harris. A new tool for measuring student behavioral engagement in large university classes. *Journal of College Science Teaching*, 44(6):83–91, 2015.

[25] H. Lei, Y. Cui, and W. Zhou. Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: an international journal*, 46:517–528, 03 2018. doi: 10.2224/sbp.7054.

[26] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[27] J. Li, D. Li, S. Savarese, and S. 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.

[28] L. Li, Z. Xie, M. Li, S. Chen, P. Wang, L. Chen, Y. Yang, B. Wang, and L. Kong. Silkie: Preference distillation for large visual language models, 2023.

[29] S. Li, R. Lin, and S. Pei. Multi-modal preference alignment remedies regression of visual instruction tuning on language model, 2024.

[30] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

[31] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[32] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022.

[33] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[34] D. McDuff, R. el Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected "in-the-wild". In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 881–888, 2013. doi: 10.1109/CVPRW.2013.130.

[35] R. Mokady, A. Hertz, and A. H. Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[37] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.

[38] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[39] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023.

[40] D. R.-D. Robert J. Jagers and B. Williams. Transformative social and emotional learning (sel): Toward sel in service of educational equity and excellence. *Educational Psychologist*, 54(3):162–184, 2019. doi: 10.1080/00461520.2019.1623032. URL https://doi.org/10.1080/00461520.2019.1623032.

[41] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[42] P. Sharma, S. Joshi, S. Gautam, S. Maharjan, S. R. Khanal, M. C. Reis, J. Barroso, and V. M. de Jesus Filipe. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning, 2023.

[43] C. Thomas and D. B. Jayagopi. Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, MIE 2017, page 33–40, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450355575. doi: 10.1145/3139513.3139514. URL https://doi.org/10.1145/3139513.3139514.

[44] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[45] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[46] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[47] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[48] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia. Aff-wild: Valence and arousal 'in-the-wild' challenge. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1980–1987, 2017. doi: 10.1109/CVPRW.2017.248.

[49] H. Zhang, X. Li, and L. Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[50] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[51] Z. Zhao, B. Wang, L. Ouyang, X. Dong, J. Wang, and C. He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization, 2024.

[52] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.