# Integrating Physiological Signals in Multimodal Large Language Models for Estimating Human Affective States

Pavana Pradeep, Stephanie Millwood, Krishna Kant, Longin Jan Latecki
Temple University

## Abstract

*We explore effective means to enhance estimation of human affective states by integrating video analysis with physiological signals representing the heart and breathing rate. These signals are derived from the same videos (i.e., no additional input is needed) thanks to the recent progress in remote photoplethysmography (rPPG). Our goal is to integrate rPPG derived signals into foundation vision-language models (VLMs) since they have great ability of understanding videos due to their pretraining on a huge number of images and videos. However, we face the problem that VLMs have never been exposed to physiological signals. To address this issue, we convert rPPG derived signals to images based on Toeplitz matrix transformation and demonstrate that this transformation outperforms other representations of physiological signals. We call our model Physio-VLM (PVLM). Using publicly available datasets, we show that PVLM outperforms state-of-the-art (S0TA) methods by a huge margin in estimating human affective states like stress, frustration, confusion, and boredom, e.g., we increase the classification accuracy of four levels of boredom by over 50% (from 36% to over 90%) on the DAiSEE dataset. Moreover, PVLM improves the accuracy of student engagement prediction by over 20%. We will release the code and our versions of the datasets upon publication.*

## 1. Introduction

As Vision Large Language Models (VLMs) enter an increasing number of areas in our lives, we face the challenge of how to add new modalities to VLMs on which VLMs were not pretrained. Here, we focus on physiological signals that are invariably pseudo-periodic, but the challenge also applies to other modalities. While it is easy to convert new types of data to Large Language Model (LLM) tokens, pretrained foundation VLMs or Multimodal LLMs (MLLMs) may never have been exposed to such tokens during their training process. The goal of training foundation models to understand additional modalities remains out of reach for many research groups due to high energy usage and cost.

To address this problem, we propose a model called *Physio-VLM* or PVLM, in which we integrate physiological signals into a VLM by converting them to images in a way that retains the periodicity information. According to [33], periodicity is an essential aspect of the relationship between photoplethysmography (PPG) signal and mental effort. Since for physiological signals, it is important to retain the pseudo-periodic nature when converting from 1D to image format, we use the Toeplitz matrix for the conversion, which essentially repeats the signal in successive rows with shifts so that the periodicity can be identified visually in the representation. For comparison, we also consider other 2D representations of 1D signals, such as Discrete Wavelet Transform (DWT) and Short-Time Fourier Transform (STFT) images. Our experimental results demonstrate that augmenting video frames with Toeplitz images yields the highest performance gains in various tasks related to remote estimation of human engagement and other affective states. A human's affective state encompasses their internal experience of emotions, moods, and feelings, influencing their engagement levels in learning. It spans a wide range of emotional conditions such as joy, frustration, or boredom, which can impact motivation and participation.We also demonstrate that Toeplitz transformation significantly outperforms direct conversion of PPG to LLM tokens.

We finetune a VLM with the standard loss of next-token prediction. However, we stress that the next token prediction applied to only video input is not able to extract the PPG information included in videos. As demonstrated in our experimental results, since explicitly adding the remote PPG (rPPG) signal and its filtered version significantly increases the performance. The rPPG signals are extracted using a frozen vision transformer RhythmFormer [37], which was trained by minimizing the difference between the predicted rPPG and direct contact PPG signals.

Our main contributions can be summarized as:

- An observation that explicit integration of the physiological signals can significantly enhance the accuracy of estimating engagement and other affective states from face videos. Interestingly, the signals are derived from the

same videos, thanks to the recent progress in remote photoplethysmography (rPPG) [37]. Our work also demonstrates that VLMs alone are unable to extract the physiological signals directly from videos. This is most likely due to the VLM's loss function, which is based on the next token likelihood of LLMs. So, creating an explicit and adequate representation of physiological signals for VLMs is an important contribution of our work.

- Answering this question: How to make physiological signals understandable for foundation VLMs under the following constraints: (1) VLMs were not trained/exposed to such signals, and (2) we have a very small amount of data for finetuning. The key to our answer is transforming 1D physiological signals into specially constructed 2D Toeplitz images, which benefits from VLM's familiarity with images. We also show that converting physiological signals to Toeplitz images for integration with VLMs provides better and more robust results than other techniques like DWT and STFT transformations.
- The use of emerging VLMs provides a more straightforward and generalizable method for the estimation of affective states than traditional vision methods focused on individual tasks such as facial expression or posture recognition.

It is important to note that the finetuned PVLM model can be run locally on each participant's computer, with only the high-level feedback transmitted to other parties (e.g., the instructor in an online education setting), i.e., without transmitting the participant's raw video feed to any other party.

The rest of the paper is organized as follows. Section 2 introduces the problem of student engagement and behavior markers and the role of PPG and other measures in engagement. Section 3 describes how we extend the VLM architecture to include the rPPG. Section 4 explains the experimental setup, compares our method to other methods, and analyzes the results. Finally, section 5 concludes the paper.

## 2. Background

### 2.1. Online Attentiveness Monitoring

In the wake of the COVID-19 pandemic, online engagements, including education and business meetings, have become ubiquitous and are expected to remain entrenched due to convenience and lower costs. However, online learning makes it difficult for an instructor to make face-to-face contact and assess the engagement level of participants [31]. It has been shown that engagement in virtual classrooms positively correlates with academic achievement [19], and it is expected that better engagement would lead to better outcomes in other settings as well. Thus, it is highly desirable to measure engagement automatically and without the need for any contact devices. The camera and microphone, being essential for online participation, provide natural and non-intrusive modalities for monitoring.

Although participant's speaking behavior is a useful modality in interactive business meetings, it is largely absent in lecture settings. We thus largely focus on video alone in this paper. Videos can be used to monitor many behavioral aspects such as eye gaze, facial expression, and posture, but these may be inadequate. Engagement or lack thereof is often reflected in the physiological parameters such as heart rate, blood pressure, skin resistance, etc. [10, 31]. These can be estimated from the chromatic variations in the skin as a function of physiological parameters and thus estimable remotely from changes over successive video frames, known as remote Photoplethysmography (rPPG). Clearly, this information is already a part of the recorded videos; however, it turns out that extracting it explicitly and then integrating it with normal video processing can significantly enhance the attentiveness detection.

While traditional deep learning models can be used for visual behavior recognition, they have significant limitations, including the need to devise specialized algorithms and collect substantial training data. In contrast, the emerging VLMs can be finetuned more efficiently and have better generalizability. However, since pretrained VLMs were most likely not exposed to rPPG signals, we need to integrate them into the VLMs explicitly.

### 2.2. Engagement Markers

Although people intuitively understand the concept of engagement, it is tough to define it objectively, as it depends on numerous factors that are not well understood. In the context of determining engagement for more effective teaching in a virtual environment, the goal is to determine engagement level non-intrusively (i.e., largely through video monitoring means) and if this determination is at least as accurately as done by the instructors in a physical class, the technique would already be helpful for online education. Accurately estimating some recognizable physiological and emotional aspects of the engagement can be highly valuable for online teaching [31]. Sharma et al. [28] fuse eye and head movement information with facial emotions to generate an engagement index. Huang et al. [16] also used features such as eye gaze direction, head pose, and eye coordinates and achieved higher accuracy than using these features individually.

Several works have attempted to recognize emotions from facial expressions and then translate them to engagement measures [8, 13, 32]. Facial expression datasets focused on facial expressions and head orientations often need more integration with online learning environments or the context created by teacher instructions. Datasets that do consider the student learning environment include the Student Engagement Dataset (SED) [8], DAiSEE [13], Enga-

geNet [29], WACV dataset [5], and UBFC-Phys [27] (Video dataset with ground truth physiological measurements).

### 2.3. Photoplethysmography and Engagement

Photoplethysmography (PPG) is a non-invasive, contact optical technique to monitor heart rate but can also provide insights into other physiological parameters. The PPG measurements can reveal several quantities, including (a) Blood Volume Pulse (BVP) – the amount of blood being pumped with each heart-beat; (b) Heart rate (HR) and heart rate variability (HRV); and (c) energy in various frequency bands [33]. BVP, HR, and HRV can be considered critical properties of the PPG signal [23].

The low-frequency part of the PPG spectra (below 0.5 Hz) primarily represents phenomena unrelated to the heart, such as movement, breathing, and thermoregulation. Of these, the range of 0.1-0.4Hz can be considered as corresponding to breathing [14], although motion could also occupy this range.

It has been shown in [26] that mental focus reduces the amplitude of the PPG signal. Another aspect concerns heart rate variability (HRV) and width of the pulses. HRV is known to increase with the engagement level [27], and pulse width is known to decrease [26]. [22] shows that HRV is sensitive to variations in mental effort; as cognitive effort increases, the power decreases in the lower frequency band 0.04–0.15 Hz. Many other works have proposed specialized models to analyze PPG signal, such as [5, 18, 21].

Given the basic principles of PPG, the so-called remote PPG (rPPG) attempts to estimate either the chromatic signals or, more directly, the heart rate from regular images. As expected, the lighting variations due to ambient lighting differences, subject posture, skin type, movement, etc., make rPPG much more noisy [22] than contact PPG signal. Furthermore, the rPPG signal is available only from the exposed skin and varies significantly from region to region. Ref [6] tries to select the regions of interest in an unsupervised manner to estimate PPG. [27] compares ground truth PPG signal against rPPG measurement and finds 85% accuracy. Ref [30] discusses using traditional CNN to estimate heart rate from a sequence of video images. [36] uses a similar 3D-CNN method. [34] discusses a method to generate face videos with embedded PPG signals synthetically. It generates realistic 3D face models and infuses them with PPG signals derived from a ground-truth pulse oximeter dataset.

Two recent methods, RhythmMamba [38] and RhythmFormer [37], can generate high-quality rPPG signals from face videos with RMSE of only a few percent with respect to ground truth PPG signals. The former is based on Mamba (or state space model), whereas the latter is based on transformers. We found that RhythmFormer works substantially better than RhythmMamba, and we have used it for the purposes of this paper. Overall, we find that the rPPG signal can be estimated reasonably well, often better than the PPG signal [27]. The primary reason for it is the ability to detect the most promising exposed skin regions (e.g., face and arms) rather than measuring them at a single spot. This raises the intriguing possibility of using the rPPG signal to enhance the video analysis as it relates to predicting the attention level in online engagements, which we explore in the following parts.

Given a normal rPPG frequency of around 1 Hz but a 30 FPS rate for videos, the PPG signal usually corresponds to all videos or video chunks rather than individual frames. For example, for a 10-second video, RhythmFormer emits a $1 \times 296$ vector of PPG values, or 296 samples per 10 seconds, which, according to Nyquist criteria, can reconstruct signals up to 8 Hz, or the 4th harmonic of a 2 Hz (or 120 beats/minute) heart rate.

## 3. Methodology for rPPG Integration

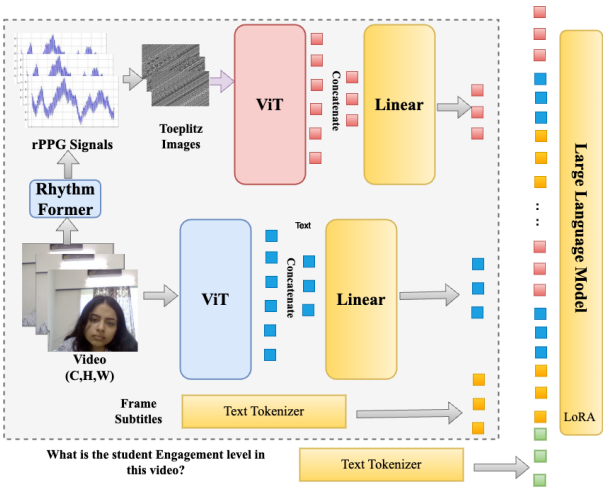



Figure 1. Proposed rPPG Integration Architecture.

### 3.1. Converting 1D Signals to Images

There are many methods in the literature for converting time-series to images as discussed in [12]. Some of these, such as Gramian Angular Field (GAF) which represents temporal changes using angular coordinates, are designed for arbitrary signals, whereas we need to focus on those that consider the pseudo-periodicity and other unique characteristics of the rPPG signals. The most general way to deal with spatio-temporal aspects of the signals is by using the time-frequency field represented (a) Spectrogram, produced by Short-time Fourier Transform (STFT), or (b) Scalogram, produced by the (Discrete) Wavelet Transform (DWT). Of these, DWT is more general and revealing in that it provides a true trade-off between frequency and time, whereas

STFT chooses a constant time-window to extract the frequency behavior. Neither is focused on the periodicity of the signal; instead, they are designed to capture all significant frequency components of the signal.

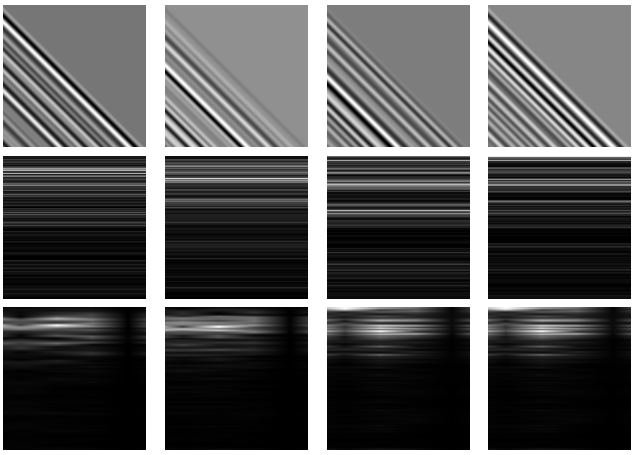
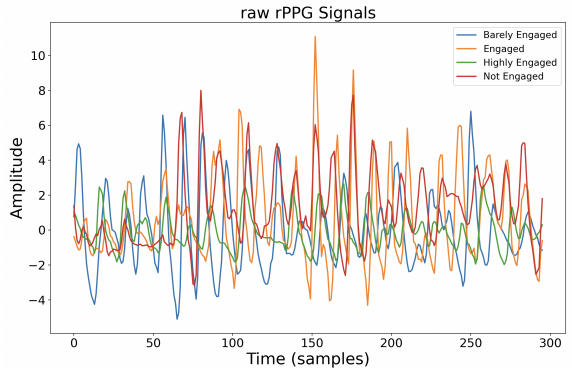


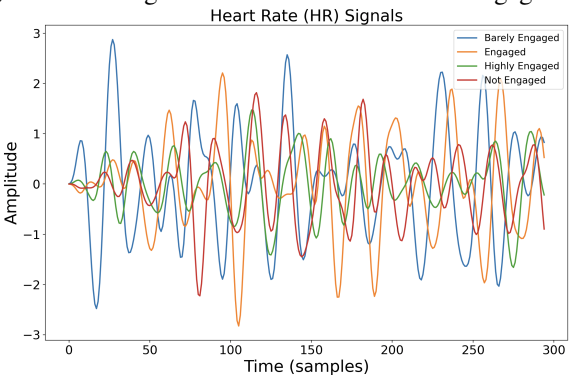
Figure 2. Image representations of the HR signals in Fig. 3(b). Row 1: Toeplitz, Row 2: DWT, Row 3: STFT images.

We utilize the 2D Toeplitz matrix representation to convert rPPG signals to images. It repeats the 1D signal with increasing shift to capture the periodicity directly. In our version of the Toeplitz matrix, a 1D signal $[s_0, s_1, ..., s_{2(n-1)}]$ is represented by a square matrix of size $n \times n$ whose $m$th row ($m = 0..n-1$) is $[s_m, s_{m+1}, ..., s_{m+n-1}]$. We can represent this matrix as a gray-level image, which we call **Teoplitz image**. Fig. 2 presents the Toeplitz images generated from HR signals, corresponding to each of the four affective states in the SED dataset as shown in Fig. 3. Corresponding sample screenshots are shown in Fig. 4. The slanting streaks in the Toeplitz images correspond to the periodic behavior of the signal. The lines correspond to individual peaks, and the spacing between them corresponds to the width. Since both STFT and DWT capture spectral characteristics across time, those representations show horizontal lines because of a lack of significant variability over time. Also, DWT representation is more varied as it considers a larger range of scales than STFT, which is limited to short time windows.

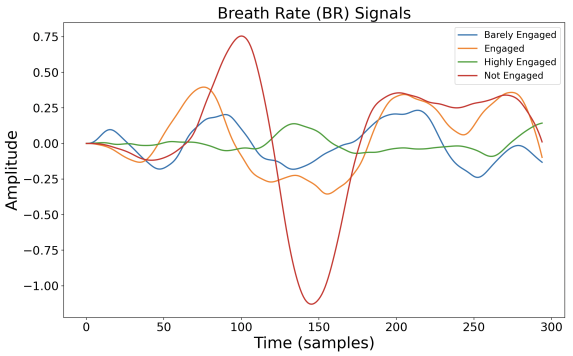
Since VLMs are trained on millions of natural images, it may be also easier for VLMs to perceive the distinctive pattern in Toeplitz image as compared with STFT or DWT images. This is also confirmed by our experimental results shown in Sec. 4.3. The benefits of Toeplitz images for multimodal emotion recognition are also demonstrated in [3], where they are used as input to a CNN. However, to the best of our knowledge, we are the first to demonstrate that Toeplitz images effectively represent 1D rPPG signals for VLMs.




(a) raw rPPG Signals for each level of student engagement




(b) HR Signals for each level of student engagement




(c) BR Signals for each level of student engagement

Figure 3. rPPG signals for each level of student engagement in SED [8] Dataset.

### 3.2. Filtering rPPG Signal in Frequency Domain

The rPPG signal contains both valuable information and low-frequency noise resulting from motion and lighting variations since PPG measurements are affected significantly by the lighting conditions and light absorption by the skin [9, 24]. The main frequency of heartbeat is around 1 Hz (60 beats/min), but due to the nonsinusoidal shape of the signal, one sees up to 3 or 4 harmonics going up to $\sim$6 Hz [23]. Some high frequency noise can also exist due to irregularities in the heart operation. Filtering methods have been proposed in the literature for PPG [25] and also for

(a) highly-engaged (b) barely-engaged

(c) engaged (d) not-engaged

Figure 4. Screenshots representing videos from SED [8] corresponding to rPPG signals in Fig. 3.

the underlying ECG signal [35]. According to our experiments, simple frequency based filtering seems adequate. In particular, one could remove both the low-frequency noise (motion, breathing, etc.) and high-frequency components by considering the signal in the range of [0.5-1.6] Hz. This range corresponds to the systolic peak of heartbeat [23, 25], which signifies the oxygenated blood being pushed through the arteries. Therefore, we call the obtained signal the **HR signal**.

As stated above, HRV is sensitive to variations in mental effort, and changes in cognitive effort can be detected in the lower frequency band 0.04–0.15 Hz [22]. Ref[33] analyzes the relationship between large numbers of PPG measures and the mental effort required for tasks. The results indicate that only a few measures are statistically significant.

As our experimental results demonstrate, the proposed method for modeling affective states can be further enhanced by incorporating breathing signals. The breathing signals reside within the low-frequency range of PPG/ECG signals, and numerous methods have been devised to extract them [7] so as to minimize the impact of motion, thermoregulation and other types of low frequency noise. For our purposes, a simple bandpass filter in the respiratory band seems adequate to extract the breathing rate called **BR signal**. The extremes of respiratory range can extend to 6 to 33 breaths/min [2], but for rPPG signals extracted from face videos of online education/meeting type of environments, a range of 0.15 Hz to 0.4 Hz seems quite adequate, with a peak typically around 18 breaths/min (0.3 Hz) [7].

Finally, we convert the HR and BR signal to Toeplitz images as described in Sec. 3.1.

### 3.3. Integrating rPPG Signal with VLM based engagement level estimation

For the integration of rPPG signals, we choose an existing multimodal LLM, namely, MiniGPT4-video [4]. The MiniGPT4-video architecture effectively tackles visual question answering by integrating visual and conversational comprehension in the video domain across multiple frames. Fig. 1 shows the overall architecture. MiniGPT4-video extracts two components: (1) visual tokens from each video frame using EVA-ViT-G [11], and then concatenates each adjacent visual token into a single token, followed by mapping into the large language model space using a linear layer, and (2) language tokens of subtitles which correspond to each video frame from the LLM tokenizer.

Due to limitations imposed by the context window of the LLM, each video is subjected to frame sub-sampling in the MiniGPT4-video framework. Since we use MiniGPT4-video with Mistral [17] as the LLM, the maximal number of images that can be processed for a given video clip is set to 90. In our setting, we have two sources of images: subsampled video frames and Toeplitz images. We experimentally determined that the equal ratio of 45 subsampled video frames and 45 Toeplitz images performs well. For example, we noticed a decline in accuracy when we allocated 60 images to the subsampled video frames and 30 to Toeplitz images. This is an interesting observation since we have only one unique Toeplitz image per video clip, which means that we repeat it 45 times for each video clip while each of the 45 video frames is different. For the original verison of MiniGPT4-video, all 90 images are subsampled video frames for each clip.

When using both HR and BR signals, we have three sources of images: subsampled video frames, HR Toeplitz images, and BR Toeplitz images. We experimentally established that the combination of 45 subsampled video frames, 25 HR Toeplitz images, and 20 BR Toeplitz images yields excellent performance.

## 4. Experimental Evaluation

We assess the estimation of engagement levels using the EngageNet [29] and Student Engagement Dataset (SED) [8] with and without the addition of physiological signals as Toeplitz images. As we clearly demonstrate, inclusion of heart rate (HR) and breath rate (BR) signals, both derived from rPPG signals, significantly enhances accuracy. In addition to estimating the student engagement, we also apply our method to stress estimation on UBFC-PHYS Dataset [27] as well as DAiSEE [13] dataset to demonstrate that our approach can effectively model emotional states. DAiSEE contains four levels of emotional states for boredom, confusion, engagement, and frustration. The datasets are described in Sec. 4.4.

### 4.1. Evaluation Metrics

In addition to standard classification accuracy, we employed GPT-3.5 turbo to juxtapose model outputs with ground truth data, computing scores aligned with MiniGPT4-video evaluation methodology, described in [20]. These scores range from 0 to 5 (perfect) and measure how well the output produced corresponds to the semantic meaning of the class designation. The scores include correctness, consistency across different responses, contextual understanding, and understanding of temporal changes. Only the first three are relevant to our case.

Further, we used the Langchain wrapper library [1] to determine the accuracy of predicted labels. LangChain simplifies converting model output to JSON format by organizing the model's predictions and responses into key-value pairs. LangChain applies its pre-configured structured output parser to ensure the data is formatted into a structured JSON after receiving a raw output from our fine-tuned model. When the output is converted to JSON format, the relevant information, including predicted class labels, can be extracted from the JSON object. Once the predicted label is extracted from the JSON output, LangChain uses string-matching techniques to compare it with the actual class label. It ensures that the predicted text matches approximately (using a fuzzy matching method, which uses a Levenshtein (or edit) distance measure to handle minor variations or typos) with the actual label. Then, the model's accuracy is calculated by dividing the number of correct predictions by the total number of predictions, yielding a quantitative measure of model performance.

### 4.2. Ablation Studies and Comparison to SOTA

Our primary baseline is MiniGPT4-video without any rPPG signals, denoted as VLM w/o rPPG. It is fine-tuned and evaluated on the same datasets as the proposed PVLM. All physiological signals, raw rPPG (no bandpass filtering), HR, and BR are incorporated into MiniGPT4-video as Toeplitz images unless explicitly stated otherwise. Therefore, VLM with any of these signals is called PVLM. Our main contribution is PVLM with HR+BR.

Table 1. Results on EngageNet [29]

| Performance | VLM w/o rPPG | PVLM with raw rPPG | PVLM with HR | PVLM with HR+BR |
|---|---|---|---|---|
| Overall Accuracy (%) | 78.12 | 80.03 | 85.25 | **89.24** |
| Correctness | 3.78 | 3.98 | 4.25 | **4.47** |
| Contextual understanding | 2.05 | 3.01 | 4.07 | **4.31** |
| Consistency | 3.62 | 4.01 | 4.11 | **4.38** |

Tables 1 and 2 demonstrate the benefits of adding physiological signals as raw rPPG, HR, and HR+BR. The performance is measured in terms of classification accuracy and GPT-3.5-based evaluation metrics as defined in Section 4.1. First, we observe that a higher overall accuracy value correlates to higher scores of GPT-3.5 based evaluation metrics.

Table 2. Results on SED (trained on EngageNet) [8]

| Performance | VLM w/o rPPG | PVLM with raw rPPG | PVLM with HR | PVLM with HR+BR |
|---|---|---|---|---|
| Overall Accuracy (%) | 77.41 | 79.81 | 84.61 | **88.64** |
| Correctness | 3.56 | 3.91 | 4.17 | **4.32** |
| Contextual understanding | 2.04 | 3.11 | 4.11 | **4.23** |
| Temporal understanding | 2.11 | 2.81 | 4.02 | **4.21** |

Regarding the performance, adding HR and BR as Toeplitz improves by over 11% the classification accuracy of MiniGPT4-video finetuned on the same datasets without any explicit rPPG signal. So, finetuning with the rPPG derived signals is essential in estimating engagement levels, which also demonstrates that the VLM is unable to extract physiological information with the standard next-token prediction loss.

It can be seen that accuracy improves with the use of raw rPPG but only by 2%. We used the signal as estimated by RhythmFormer without any frequency band filtering, and hence the designation "raw rPPG". As demonstrated in the next two columns, the frequency band filtering is essential when dealing with rPPG signals. It is seen that the HR signal improves the results by 7%. (frequency range of [0.5-1.6] Hz, which corresponds to the heart rate signal). Finally, the last column shows the result of adding the low-frequency component (in the range of [0.15-0.4] Hz), which corresponds to the breath rate signal. This improves the results by another 4% increase in accuracy compared to the use of HR alone.

The authors of EngageNet [29] did not release their test dataset, but they reported the results of their best-performing method (Transformer Fusion Model) on their validation dataset, which is the dataset we used in all our tests in Table 1. The accuracy reported in [29] is 68.49%. So, the proposed PVLM with HR+BR improves it by **over 20%** over the Transformer Fusion Model in [29].

Table 2 evaluates the VLMs trained on Engagenet training dataset on SED dataset. The results indicate strong generalization, as the model performs only slightly worse on SED compared to Engagenet. We observe that SED is created with a different goal, by a different group of researchers, and for a different purpose, as discussed in Sec. 4.4. So, its videos can be viewed as out-of-distribution samples for the proposed approach.

Table 3. Comparison to SOTA as accuracy (%) on DAiSEE [13]

| Method | Attentive State Type | | | |
|---|---|---|---|---|
| | Engagement | Boredom | Confusion | Frustration |
| EmotionNet [13] | 51.07 | 35.89 | 57.45 | 73.09 |
| VLM w/o rPPG | 73.25 | 69.81 | 74.15 | 78.42 |
| PVLM with HR+BR | **89.64** | **90.75** | **91.15** | **91.75** |

Table 3 compares the proposed method to EmotionNet [13] on estimating various human affective states: boredom, confusion, and frustration, besides assessing participant engagement, on the DAiSEE [13] dataset. We observe

a huge performance improvement in classification accuracy in comparison to the original method EmotionNet and MiniGPT4-video. In particular, PVLM with HR+BR improves **over 50%** the classification accuracy of boredom in comparison to EmotionNet, and it improves over 20% over the MiniGPT4-video. The remarkable improvement over the MiniGPT4-video baseline for all four affective states clearly demonstrates the benefits of the proposed integration of HR and BR signals as Toeplitz images in the VLM framework. The results of this section thus demonstrate that our proposed method can be utilized to assess a variety of affective states.

Table 4 presents the results of our PVLM with HR+BR on differentiating between stress and non-stress states on the UBFC-PHYS dataset. Several SOTA models were evaluated on this dataset, and as reported in Table 4 of [27], the best performing model achieved the classification accuracy of 85.48%. So, the proposed PVLM surpasses the classification accuracy of all methods referenced in [27] by 4%.

Table 4. Results on UBFC-PHYS [27]

| Performance | PVLM with contact PPG | PVLM with raw rPPG | PVLM with HR+BR |
|---|---|---|---|
| Overall Accuracy (%) | 80.45 | 82.45 | **89.4** |
| Correctness | 4.06 | 4.10 | **4.52** |
| Contextual understanding | 3.63 | 3.72 | **4.45** |
| Consistency | 3.82 | 3.96 | **4.53** |

The UBFC-PHYS dataset also includes contact-PPG pulse signals captured from participants utilizing an E4 bracelet equipped with an accelerometer, which calculates the Inter-Beat Intervals from the Blood Volume Pulse signal, forming the contact-PPG signals. So, here we are able to compare contact-PPG and rPPG signals both represented as Toeplitz images. Interestingly, PVLM with rPPG yields better results than with contact PPG. As in other experiments, we utilized RhythmFormer [37] to extract rPPG signals from videos. So, this result may indicate that rPPG may be able to capture additional physiological markers not contained in PPG signals.

### 4.3. Toeplitz vs. with STFT and DWT images

Table 5. Results of PVLM with HR + BR, STFT and DWT images

| Method | Accuracy (%) | | | |
|---|---|---|---|---|
| | EngageNet | SED | UBFC-PHYS | DAiSEE |
| PVLM with HR+BR | **89.24** | **88.64** | **89.4** | **91.75** |
| STFT with HR + BR | 85.15 | 83.76 | 85.8 | 86.12 |
| DWT with HR + BR | 83.84 | 83.05 | 84.65 | 85.23 |

We compare Toeplitz image representation of HR and BR, both derived from the rPPG signal, to those of STFT and DWT images in Table 5, i.e., when Toeplitz images have been replaced with STFT and DWT images, respectively. The superior result of the proposed Toeplitz images (increase by 4 to 5%) confirms its expected superiority over DTW and STFT for representing pseudo-periodic, physiological signals as input to VLMs.

More experimental details are provided in the supplementary material.

### 4.4. Details on Datasets Used

EngageNet has around 30 hours of video data of over 100 subjects whose engagement is classified into four classes: highly-engaged, engaged, barely-engaged, and not-engaged. The videos were recorded on subjects' devices using the cameras that they would use in an actual online class, such as their built-in laptop webcams. Since we were unable to obtain the test set, we used the authors' validation set of 11 participants as our test set. We have a total of 7879 videos in the training set and a total of 975 videos in the test set. The class distribution of the training set is as follows: not-engaged: 1446 clips; barely-engaged: 1035 clips; engaged: 1658 clips; highly-engaged: 3740 clips. The class distribution of the validation set is as follows: not-engaged: 108 clips; barely-engaged: 71 clips; engaged: 249 clips; highly-engaged: 547 clips.

The SED is similar to the EngageNet dataset in that it consists of videos of students in a similar age range completing educational computer-based tasks in front of a webcam. However, there are several key differences between the datasets. A major difference between the two datasets is that unlike in the EngageNet study, the online lessons used for the SED data collection required students to spend some of their time doing work on paper, i.e., the subjects were often not looking at the screen, but instead looking down and in front of or to the side of the screen. This allowed us to evaluate our model's ability to classify engagement with a wider variety of activity types. This is an essential consideration since online classes often include a mixture of lectures and individual work. Since SED videos are of different lengths, we chunked them into clips of 10 seconds in length.

We have relabeled SED videos to match the labels of EngageNet. Since the relabeling step was performed automatically, we ended up with 170 video clips that could be consistently relabeled. Their class distribution is as follows: not-engaged: 18 clips; barely-engaged: 41 clips; engaged: 53 clips; highly-engaged: 58 clips.

UBFC-PHYS [27] is a multimodal dataset concerning the impact of social stress on both in-person and remote physiological responses. The dataset includes a collection of electrodermal activity, blood volume pulse signals (contact-PPG), and videos from 56 participants. The video-recording session involved three tasks: a 10-minute rest task named T1, a 3–4 minute speech task, T2, and a 6-minute arithmetic task, T3. In the study, participants completed three tasks: a rest task, a speech task, and an arithmetic task. During the rest task, participants stayed silent. For the

speech task, they were randomly assigned to either a harder "test" scenario (simulated job interview) or an easier "ctrl" scenario (recalling a positive holiday or dream vacation). In the arithmetic task, they either counted down from 2023 by 17s (test) or from 2025 by 10s (ctrl). All tasks were categorized as non-stress (T1) or stress (T2, T3) states, with stress levels inferred from T2 and T3 for subjects with valid signals. Only PRV features were used to define the stress state. For experiments, each task's video was edited to three minutes, and we used binary classification: stress (56 clips) and non-stress (74 clips).

We also utilize the DAiSEE dataset [13] to demonstrate the applicability of the proposed approach to a wide range of human affective states. It has 9068 videos of 112 subjects (80 males and 32 females). Each video clip is labeled according to four different affective states: engagement, boredom, confusion, and frustration. For each state, the clip is classified into one of four levels: very low, low, high, or very high. The labels for each affective state are independent of one another, so a clip can have any combination of levels for the four states. For example, one clip could have a label of very high engagement, low boredom, very low confusion, and low frustration.

### 4.5. Fine-tuning Pipeline

We finetune MiniGPT4-video [4] with the standard loss of next token prediction. In the fine-tuning stage, we train a linear layer, which projects both the Toeplitz image features and the visual features encoded by the vision encoder to the LLM's text space with captioning loss. (We use Mistral [17] as LLM in our experiments.) During this stage, we use the predefined prompts in the following template:
[INST] <img><FrameFeature_1><rppg><Rppg_Feature_1><Sub> <Subtitletext_1><img><FrameFeature_2><rppg><Rppg_Feature_2> <Sub><Subtitletext_2> ... <img><FrameFeature_N><rppg> <Rppg_Feature_N><Sub><Subtitletext_N> [/INST]{prompt here}

In this prompt, each <FrameFeature> is replaced by the sampled video frame encoded by the visual backbone, and each <Rppg_Feature> is replaced by the Toeplitz image encoded by another visual backbone. The <Subtitletext> represents the subtitle for the corresponding frame. It provides a more detailed description of a given affective state, which is another reason why VLMs outperform vision-only models for classification tasks. We used subtitles to define the engagement categories in the training videos, as the authors of MiniGPT4-video evaluated their model with and without subtitles. Their findings indicate a notable improvement in accuracy when subtitles are used, particularly for the TVQA dataset (please refer to Table 2 in [4]). The subtitles were left empty for evaluation/test videos.

The </INST> represents a randomly sampled instruction from our predefined instruction set containing variant forms of instruction, which are as follows for EnagageNet and SED datasets: (1) What is the current level of student engagement in this video?, (2) How actively is the student engaged in this video?, (3) How to assess student engagement in this video?, and (4) How engaged are students in this video?. For each sampled frame, the extracted visual and text tokens representing the subtitles are concatenated. Instruction tokens are appended to the end of the input sequence, and the model then generates the answer to the question.

All experiments were performed on a server with two NVIDIA RTX A6000 GPUs, each equipped with 10752 CUDA cores and 48GB GDRR6 memory. During the fine-tuning stage, we consistently used a batch size of 1 and employed the AdamW optimizer with a cosine learning rate scheduler set to 1e4. We utilized LoRA [15] to fine-tune the modified MiniGPT4-video and conduct efficient multi-modal language model fine-tuning. In particular, we optimized the $W_q$ and $W_v$ components by adjusting their rank (r) to 64 and setting the LoRA-alpha constant to 16. Throughout our fine-tuning stage, we keep a consistent resolution of $224 \times 224$ pixels for both video frame and Toeplitz images to ensure uniformity across all stages.

## 5. Conclusions and Future Work

In this paper, we explore multiple ways of integrating the video-derived rPPG or remote PPG, signal into the existing VLM pipelines. We further explore the heart and breath rate signals extracted from rPPG to enhance the accuracy of student engagement level estimation in virtual learning environments as well as other affective states. We show that converting 1D rPPG signals to 2D Toeplitz images enables LLMs to understand them after finetuning on relatively small datasets. We also show that if the rPPG signal is preprocessed to remove much of the considerable amount of noise that such a signal carries, the integration is more beneficial and improves the accuracy of the estimation. As our experimental results demonstrate, PVLM with HR and BR achieves double digit improvements over SOTA methods and over MiniGPT4-video without any physiological signals. The achieved accuracy of around 90% in estimating various human affective states demonstrates the potential impact of the proposed method on many applications ranging from remote learning to human-robot interaction. The strong performance of our model when trained on EngageNet and evaluated on SED shows the generalizability of the method. An intriguing question is whether the proposed Toeplitz image representation is also suitable for representing other signals like EEG.

## References


[1] Langchain framework. https://js.langchain.com/v0.1/docs/get_started/introduction. Accessed: 2024-08-01. 6

[2] Ali Adami, Reza Boostani, Faezeh Marzbanrad, and Peter H Charlton. A new framework to estimate breathing rate from electrocardiogram, photoplethysmogram, and blood pressure signals. *IEEE Access*, 9:45832–45844, 2021. 5

[3] Kamran Ali and Charles E Hughes. A unified transformer-based network for multimodal emotion recognition. *arXiv preprint arXiv:2308.14160*, 2023. 4

[4] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024. 5, 8

[5] Sarthak Batra, Hewei Wang, Avishek Nag, Philippe Brodeur, Marianne Checkley, Annette Klinkert, and Soumyabrata Dev. Dmcnet: Diversified model combination network for understanding engagement from video screengrabs. *Systems and Soft Computing*, 4:200039, 2022. 3

[6] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 3

[7] Peter H Charlton, Drew A Birrenkott, Timothy Bonnici, Marco AF Pimentel, Alistair EW Johnson, Jordi Alastruey, Lionel Tarassenko, Peter J Watkinson, Richard Beale, and David A Clifton. Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review. *IEEE reviews in biomedical engineering*, 11:2–20, 2017. 5

[8] Kevin Delgado, Juan Manuel Origgi, Tania Hasanpoor, Hao Yu, Danielle Allessio, Ivon Arroyo, William Lee, Margrit Betke, Beverly Woolf, and Sarah Adel Bargal. Student engagement dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3628–3636, October 2021. 2, 4, 5, 6

[9] Théo Desquins, Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. A survey of photoplethysmography and imaging photoplethysmography quality assessment methods. *Applied Sciences*, 12(19):9582, 2022. 4

[10] M. Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. Engagement detection in online learning: a review. *Smart Learning Environments*. *Smart Learning Environments*, 6(1):1–20, 2019. 2

[11] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale, 2022. 5

[12] Gabriel Rodriguez Garcia, Gabriel Michau, Mélanie Ducoffe, Jayant Sen Gupta, and Olga Fink. Time series to images: Monitoring the condition of industrial assets with deep learning image processing algorithms. *arXiv preprint arXiv:2005.07031*, 2020. 3

[13] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild, 2022. 2, 5, 6, 8

[14] Judith Aa Hirsch and Beverly Bishop. Respiratory sinus arrhythmia in humans: how breathing pattern modulates heart rate. *American Journal of Physiology-Heart and Circulatory Physiology*, 241(4):H620–H629, 1981. 3

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 8

[16] Tao Huang, Yunshan Mei, Hao Zhang, Sanya Liu, and Huali 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. 2

[17] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 5, 8

[18] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. Prediction and localization of student engagement in the wild, 2018. 3

[19] Hao Lei, Yunhuo Cui, and Wenye Zhou. Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: an international journal*, 46:517–528, 03 2018. 2

[20] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424*, 2024. 6

[21] Omid Mohamad Nezami, Mark Dras, Len Hamey, Deborah Richards, Stephen Wan, and Cécile Paris. Automatic recognition of student engagement using deep learning and facial expression. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 273–289. Springer, 2020. 3

[22] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 562–576. Springer, 2019. 3, 5

[23] Christina Orphanidou. Signal quality assessment in physiological monitoring: state of the art and practical considerations. 2017. 3, 4, 5

[24] Christina Orphanidou and Christina Orphanidou. Quality assessment for the photoplethysmogram (ppg). *Signal Quality Assessment in Physiological Monitoring: State of the Art and Practical Considerations*, pages 41–63, 2018. 4

[25] Junyung Park, Hyeon Seok Seok, Sang-Su Kim, and Hangsik Shin. Photoplethysmogram analysis and applications: an integrative review. *Frontiers in Physiology*, 12:808451, 2022. 4, 5

[26] María Dolores Coca Pelaez, María Teresa Lozano Albalate, Alberto Hernando Sanz, Montserrat Aiger Vallés, and Eduardo Gil. Photoplethysmographic waveform versus heart rate variability to identify low-stress states: attention test. *IEEE journal of biomedical and health informatics*, 23(5):1940–1951, 2018. 3

[27] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 14(1):622–636, 2021. 3, 5, 7

[28] Prabin Sharma, Shubham Joshi, Subash Gautam, Sneha Maharjan, Salik Ram Khanal, Manuel Cabral Reis, João Barroso, and Vítor Manuel de Jesus Filipe. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning, 2023. 2

[29] Monisha Singh, Ximi Hoque, Donghuo Zeng, Yanan Wang, Kazushi Ikeda, and Abhinav Dhall. Do i have your attention: A large scale engagement prediction dataset and baselines. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 174–182, 2023. 3, 5, 6

[30] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, Newcastle, UK*, pages 3–6, 2018. 3

[31] Ajitha Sukumaran and Arun Manoharan. A survey on automatic engagement recognition methods: online and traditional classroom. *Indonesian Journal of Electrical Engineering and Computer Science*, 30(2):1178–1191, 2023. 2

[32] Chinchu Thomas and Dinesh Babu 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. 2

[33] Jie Wang, Tuantuan Lu, Ruogu Huang, and Yongxiang Zhao. Classifying engagement in e-learning through gru-tcn model using photoplethysmography signals. *Biomedical Signal Processing and Control*, 90:105903, 2024. 1, 3, 5

[34] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. Synthetic generation of face videos with plethysmograph physiology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20587–20596, 2022. 3

[35] Liping Xie, Zilong Li, Yihan Zhou, Yiliu He, and Jiaxin Zhu. Computational diagnostic techniques for electrocardiogram signal analysis. *Sensors*, 20(21):6318, 2020. 5

[36] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 3

[37] Bochao Zou, Zizheng Guo, Jiansheng Chen, and Huimin Ma. Rhythmformer: Extracting rppg signals based on hierarchical temporal periodic transformer. *arXiv preprint arXiv:2402.12788*, 2024. 1, 2, 3, 7

[38] Bochao Zou, Zizheng Guo, Xiaocheng Hu, and Huimin Ma. Rhythmmamba: Fast remote physiological measurement with arbitrary length videos. *arXiv preprint arXiv:2404.06483*, 2024. 3