Hierarchical Social Cue Fine-Tuning for Video-Understanding Foundation Models: A Preliminary Study with VideoLLaMA 3

Anonymous ACL submission

Abstract

High-level social cue recognition from facial data is an important but difficult task for modern computer vision systems. Foundation models (FMs) are trained on massive amounts of cross-domain data and have the potential to represent social cue concepts and the lower-level components. However, explicit bridging between levels of hierarchical social reasoning has yet to be explored through detailed fine-tuning experiments with existing open-source model architectures. We present a detailed preliminary fine-tuning study of a small, open-source video-understanding FM: VideoLLaMA 3 (seven-billion parameter version). We examined whether VideoLLaMA 3 can bridge between a low-level task (head pose classification) and a mid-level task (head gesture classification), and found that fine-tuning on head-pose and head-gesture tasks individually yielded highest macro F1 scores for each task respectively. Joint-training on both tasks yielded the second highest results, opening up future directions for exploring the link between low- and mid- level social cue tasks for FMs. Additionally, we explored the effects of different fine-tuning configurations within the original VideoLLaMA 3 training framework, and found that multi-stage instruction tuning combined with a more task-relevant focused fine-tuning, resulted in highest the model performance.

1 Introduction

Automatic social cue recognition is an important task for modern computer vision systems. Social cues, such as engagement, remain difficult for supervised computer vision systems to recognize, even though such systems have achieved strong performance on lower-level perceptual tasks (Baltrušaitis et al., 2015; Zhi et al., 2020). Foundation models (FMs) trained on large amounts of cross-domain data have the potential to represent social cue concepts and the lower-level components that give rise to them—a capability that supervised computer vision models lack, as they are constrained to task-specific labels (Brown et al., 2020).

There has been a recent increase in research on video-understanding multimodal FMs. This capability is also compute-intensive, and so accessible, high-performing video-understanding FMs still have areas for improvement (Zhang et al., 2025; Fu et al., 2025). Specifically, for higher-level social cue recognition tasks, video understanding is a key capability that computer vision systems need in order to provide real-time analysis, and that understanding is still lacking (Fu et al., 2025; Cho et al., 2025).

FMs are known to improve performance on specific downstream tasks through fine-tuning, which can require relatively less data than supervised training a classical machine learning model from scratch, with the added flexibility of being applicable to tasks with different output labels. Detailed study of fine-tuning configurations for hierarchical social cue classification tasks is still underexplored. In this work, we experimented with multiple fine-tuning configurations of a state-of-the-art small video-understanding FM, VideoLLaMA 3 (seven billion parameter model) (Zhang et al., 2025), on a low-level social-cue recognition dataset, AFLW (head pose), and on a mid-level social-cue recognition task, CCDb-HG (head gesture). We found that fine-tuning VideoLLaMA 3 on both head pose and head gesture classification improved the macro-F1 score for head pose and head gesture classification relative to the baseline VideoLLaMA 3 model. Interestingly, although head pose and head gesture tasks are related, fine-tuning on just AFLW and CCDb-HG gave the highest performance on their respective tasks. This work opens up further directions exploring the link between low and mid-level social cue recognition tasks in the context of video-understanding FMs.

2 Related Work

Head Pose Classification Head pose, such as looking up, down, left, and to the side, provides cues related to attention and intent (Murphy-Chutorian and Trivedi, 2008; Hall et al., 2019). Historically, computer vision systems used classical supervised machine learning models (Ranjan et al., 2017; Zhang et al., 2018) for head pose classification. More recently, fine-tuning with FMs for head pose estimation has been explored. For example, the vision-language model HPE-CogVLM (Tian et al., 2024) showed improvement over the current state-of-the-art convolutional neural network model used for head pose estimation.

Head Gesture Classification Head gestures, such as nodding, tilting the head, and shaking the head, convey important communicative cues in interactive settings (Vuillecard et al., 2024; Otsuka and Tsumori, 2020), and can also serve as back channel behaviors that contribute to higher-level social cues such as engagement. Supervised machine learning models, such as deep learning neural networks, have been used to classify head gesture movements (Otsuka and Tsumori, 2020; Sharma et al., 2018; Algabri et al., 2024). Recently, FMs evaluated without additional fine-tuning for head gesture classification performed worse than the previous traditional deep learning approaches (Käs et al., 2025), but fine-tuning has been shown to improve FM performance (Cho et al., 2025).

FM Fine-Tuning for Social Cue Recognition Xing et al. (2024) used an instruction-tuning approach to develop EMO-LLaMA, a multimodal FM based on LLaMA-VID (Li et al., 2024), for facial emotion recognition (FER). They fine-tuned the model using low-rank adaptation (LoRA), and incorporated facial embeddings, facial landmarks, and participant demographic information within the model architecture for video and image FER tasks. In contrast, we trained all parameters of VideoLLaMA 3 rather than employing a parameter-efficient method like LoRA to investigate the limits of fine-tuning with existing FM architectures.

Tian et al. (2024) developed HPE-CogVLM, a vision-language model for head pose estimation, and utilized a LoRA layer-based model-merging method to fine-tune a model trained on lower-level object detection tasks for head pose estimation. They found that their model-merging method avoided catastrophic forgetting in object detection.

As in both prior works, we investigated bridging between lower-level and mid-level tasks through fine-tuning. However, we specifically investigated different types of fine-tuning in VideoLLaMA 3 (Section 4) and fine-tuned all model parameters.

3 Datasets

AFLW AFLW2000-3D (Zhu et al., 2016) is a dataset comprised of facial images from the Annotated Facial Landmarks in the Wild (AFLW) dataset (Koestinger et al., 2011). We grouped images into six head pose categories based on yaw and pitch: *frontal*, *right*, *left*, *up*, *down*, and *extreme*, following Dao et al. (2024).

CCDb-HG The Cardiff Conversation Database-Head Gestures (CCDb-HG) (Vuillecard et al., 2024) contains short video clips extracted from video recordings of dyadic human conversations. The annotations are for five head gesture categories: *nod*, *shake*, *turn* , *tilt*, and *up/down*.

4 VideoLLaMA 3

VideoLLaMA 3 (Zhang et al., 2025) is a vision-centric FM trained on image and video data. Zhang et al. (2025) employed the following four training Stages in the original training for the base model: 1) vision encoder adaptation; 2) vision-language alignment; 3) multi-task fine-tuning; and 4) video-centric fine-tuning. Stages 2 and 3 were most relevant for AFLW since both take in images and focus on text-image alignment. For CCDb-HG, Stages 3 and 4 are most relevant, as they are the only Stages that train on video data. Further details of each stage can be found in Appendix A.

5 Methodology

We trained the base seven-billion parameter VideoLLaMA 3 model[1] using different training configurations described in Section 5.2. For training, we utilized two NVIDIA H100 GPUs, each with 80 GB of GPU RAM (160 GB total). For reported experiment results, we employed subject-independent five-fold cross-validation and evaluated the model based on the average macro F1 score across the five held-out test sets. Test sets were approximately 20% of the entire dataset.

[1] https://huggingface.co/DAMO-NLP-SG/VideoLLaMA3-7B

5.1 Hyperparameters

We trained VideoLLaMA 3 with default learning hyperparameters (Appendix A gives the full paramenter list). We determined feasible values for a certain set of hyperparameters by testing the highest values that were able to be stored on GPU memory for our available resources: batch size of two, 180 maximum frames, and five frames per second, and five training epochs. We used a temperature of zero (greedy decoding) for training and evaluation to test classification capabilities.

5.2 Experiments

We designed three types of experiments:

Zero-shot baseline Zero-shot evaluation of the base pre-trained VideoLLaMA 3 model on AFLW and CCDb-HG. This is the baseline for the fine-tuning experiments.

Single-task fine-tuning The base VideoLLaMA 3 model fine-tuned and evaluated on AFLW and CCDb-HG tasks, individually. Within the single-task experiments, we also tested which training Stage (Section 4) yields the highest performance. Since AFLW is an image dataset and CCDb-HG is a video dataset, we additionally tested Stages 2, 3 and Stages 3, 4 respectively.

Cross-task fine-tuning Evaluation on the mid-level CCDb-HG task of the base VideoLLaMA 3 models fine-tuned on a combination of AFLW and CCDb-HG data. This is a test of bridging between low-level head pose classification and mid-level head gesture classification. Additionally it tests whether the features learned for the low-level task improve performance on the mid-level task.

6 Results

We report average macro-F1 scores for each model training configuration in Table 1.

6.1 Zero-Shot Baselines

The pretrained VideoLLaMA 3 was evaluated in a zero-shot setting on both datasets without further fine-tuning. For head position classification, the base model achieved a mean macro F1 of 0.336, suggesting a limited ability to distinguish head orientations from images. For head gesture classification, zero-shot performance yielded a mean macro F1 score of 0.108, suggesting that temporal head gestures are a considerably harder task for FMs than static head position estimation.

6.2 Single-Task Fine-Tuning

Head Position Classification Fine-tuning VideoLLaMA 3 through Stage 2 alone achieved an average macro F1 of 0.575. Stage 3 alone achieved an average macro F1 of 0.580. Sequential, multi-stage fine-tuning through Stages 2 and 3 demonstrated the strongest results, yielding an average macro F1 score of 0.614, representing an 82.7% absolute improvement over the zero-shot baseline (0.336).

Head Gesture Classification Fine-tuning on Stage 3 alone achieved a macro F1 of 0.407. Stage 4 alone achieved 0.410. Sequential, multi-stage fine-tuning of Stages 3 and 4 yielded the highest results, with an average macro F1 score of 0.442 (309% increase over the zero-shot baseline of 0.108).

6.3 Cross-task Transfer

To investigate whether head position classification is useful for head gesture recognition, we performed cross-task fine-tuning evaluation.

We experimented with subsequent fine-tuning on CCDb-HG of the model fine-tuned on the AFLW dataset, through Stages 3 and 4 (best performing configuration in single-task fine-tuning for CCDb-HG since these Stages are able to take in video and image data). This pipeline achieved a mean macro F1 score of 0.314, a considerable improvement from the zero-shot baseline (0.108), but still lower than a model fine-tuned on only CCDb-HG (0.442).

We additionally fine-tuned VideoLLaMA 3 on combined AFLW and CCDb-HG datasets (rather than sequentially, as was done above). This model achieved an average macro F1 score of 0.605 for AFLW, and 0.406 for CCDb-HG. This was the strongest result for cross-task transfer, but still slightly lower than single-task fine-tuning (0.614 and 0.442, respectively).

7 Discussion

7.1 Task-Specific Fine-Tuning

Our results demonstrated that VideoLLaMA 3 could be effectively adapted for both head position and head gesture task-specific fine-tuning, and achieved substantial gains over the two respective zero-shot baselines. On AFLW, multi-stage fine-tuning yielded a macro F1 of 0.614 compared to a baseline of 0.336, while on CCDb-HG, multi-stage fine-tuning achieved 0.442 compared to the baseline of 0.108. These results indicate that FMs

Experiment Type	Experiment Subtype	Train Dataset	Test Dataset	F1 score
Baseline (zero-shot)	–	–	AFLW	0.336
	–	–	CCDb-HG	0.108
Single-task	Stage 2	AFLW	AFLW	0.575
	Stage 3			0.580
	Stage 2+3			**0.614**
	Stage 3	CCDb-HG	CCDb-HG	0.407
	Stage 4			0.410
	Stage 3+4			**0.442**
Cross-task	Sequential training	AFLW, CCDb	CCDb-HG	0.317
	Joint training	AFLW + CCDb	AFLW	0.605 (AFLW)
			CCDb-HG	0.406 (CCDb-HG)

Table 1: VideoLLaMA 3 model experimental results. Results are grouped into the three experiment types. Highest F1 score is in bold for evaluation on CCDb-HG test set.

can be effectively utilized for the tested tasks after specific fine-tuning.

We found that, across both datasets, multi-stage training outperformed single-stage training. There was a relatively small gap between individual Stages in terms of macro F1 performance, yet there was a substantial improvement when combining them. Stage 2 + 3 focused on vision-language alignment and multi-task fine-tuning, while Stage 3 + 4 focused on multi-task fine-tuning and video-centric fine-tuning. This result suggests that instruction tuning (multi-task fine-tuning) combined with a more task-relevant focused fine-tuning (vision-language alignment and video-centric training) is the best fine-tuning configuration for VideoLLaMA 3 for the tasks we evaluated.

7.2 Cross-Task Fine-Tuning

Both sequentially fine-tuning the model fine-tuned on AFLW, or fine-tuning on both AFLW and CCDB-HG data yielded improvement compared to the baseline for each dataset. The best results came from the model fine-tuned on both datasets. Although fine-tuning solely on each dataset resulted in the highest macro F1 scores, the improvement is small (1.49% and 8.87% improvement for AFLW and CCDb-HG) compared to the performance gap between the baseline and fine-tuning experiments. Further investigation into the link between head pose and head gesture features is required.

8 Limitations

The problem of catastrophic forgetting is often encountered when fine-tuning large language models; models improve on fine-tuned data, but worsen on previously trained data (Luo et al., 2025). We observed this in an initial experiment we performed. Although we saw better results when jointly training on both AFLW and CCDb-HG data, model-merging techniques that combine expert domain FMs for improved performance (He et al., 2025; Tian et al., 2024) are worth exploring. Furthermore, higher level tasks based on video such as head gestures, could be harder to annotate, and so another set of experiments for future work would be to trained the model on AFLW and CCDb-HG, but vary the amount of CCDb-HG data and compare that to a model not also trained on AFLW and investigate if having lower-level data for fine-tuning is beneficial in a low-data setting. Additionally, we look to study datasets other than the ones explored here, to further explore the link between tiers in the social cue task hierarchy.

9 Conclusion

We contribute a preliminary fine-tuning study of VideoLLaMA 3, with respect to hierarchical social cue reasoning. We found that fine-tuning VideoLLaMA 3 through both instruction-tuning and image/video-specific fine-tuning stages resulted in the best model performance for single-task fine-tuning experiments. We found that VideoLLaMA 3 fine-tuned on both head pose and head gesture classification yielded the second-highest model performance for head gesture classification, suggesting that low-level head pose features may be helpful for mid-level head gesture recognition, however further experiments are required to understand the extent to which head pose and head gestures features are linked.

References

Redhwan Algabri, Ahmed Abdu, and Sungon Lee. 2024. Deep learning and machine learning techniques for head pose estimation: a survey. *Artificial Intelligence Review*, 57(10):288.

Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 6, pages 1–6. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyundong Cho, Spencer Lin, Tejas Srinivasan, Michael Saxon, Deuksin Kwon, Natali T. Chavez, and Jonathan May. 2025. Can vision language models understand mimed actions? *Preprint*, arXiv:2506.21586.

Trung Tuan Dao, Duc Hong Vu, Cuong Pham, and Anh Tran. 2024. Efhq: Multi-purpose extremepose-face-hq dataset. *Preprint*, arXiv:2312.17205.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118.

Judith Hall, Terrence Horgan, and Nora Murphy. 2019. Nonverbal communication. *Annual Review of Psychology*, 70.

Yifei He, Siqi Zeng, Yuzheng Hu, Rui Yang, Tong Zhang, and Han Zhao. 2025. Mergebench: A benchmark for merging domain-specialized llms. *arXiv preprint arXiv:2505.10833*.

Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE.

Stephanie Käs, Anton Burenko, Louis Markert, Onur Alp Culha, Dennis Mack, Timm Linder, and Bastian Leibe. 2025. How do foundation models compare to skeleton-based approaches for gesture recognition in human-robot interaction? *Preprint*, arXiv:2506.20795.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*.

Erik Murphy-Chutorian and Mohan Manubhai Trivedi. 2008. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626.

Kazuhiro Otsuka and Masahiro Tsumori. 2020. Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks. *IEEE Access*, 8:217169–217195.

Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. 2017. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135.

Mohit Sharma, Dragan Ahmetovic, László A Jeni, and Kris M Kitani. 2018. Recognizing visual signatures of spontaneous head gestures. In *2018 IEEE Winter conference on applications of computer vision (WACV)*, pages 400–408. IEEE.

Yu Tian, Tianqi Shao, Tsukasa Demizu, Xuyang Wu, and Hsin-Tai Wu. 2024. Hpe-cogvlm: Advancing vision language models with a head pose grounding task. *arXiv preprint arXiv:2406.01914*.

Pierre Vuillecard, Arya Farkhondeh, Michael Villamizar, and Jean-Marc Odobez. 2024. Ccdb-hg: Novel annotations and gaze-aware representations for head gesture recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE.

Bohao Xing, Zitong Yu, Xin Liu, Kaishen Yuan, Qilang Ye, Weicheng Xie, Huanjing Yue, Jingyu Yang, and Heikki Kälviäinen. 2024. Emo-llama: Enhancing facial emotion understanding with instruction tuning. *arXiv preprint arXiv:2408.11424*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *Preprint*, arXiv:2501.13106.

Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. 2018. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3359–3368.

Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. 2020. A comprehensive survey on automatic facial action unit analysis. *The Visual Computer*, 36(5):1067–1093.

Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. 2016. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155.

A Model Training Details

A.1 VideoLLaMA 3 Training Stages

In the vision-language alignment Stage, in Stage 2, all parameters are trained in order to improve multimodal understanding. Video and image data are utilized at this Stage, and the language model, projector, and vision encoders are trained to handle multimodal data,. The multi-task fine-tuning Stage, Stage 3, primarily focuses on instruction tuning by training on a large range of multimodal query and response data. The video-centric fine-tuning Stage, Stage 4, aims to improve video understanding.

A.2 Hyperparameters

For VideoLLaMA 3, we fine-tuned the model with the following hyperparameters related to learning used by Zhang et al. (2025):

- Multimodal projector learning rate: 1^{-5}
- Vision encoder learning rate: 2^{-6}
- Large language model learning rate: 1^{-5}
- Warmup ratio: 0.03
- Image merge size: 1
- Video merge size: 2
- Model max length: 16384
- Multimodal max length: 10240

A.3 Compute Resources

For AFLW, fine-tuning for each fold in the cross-validation took approximately _ hours. For CCDb-HG, fine-tuning for each training fold took approximately six hours. We used two H100 GPUs with 80 gigabytes of GPU RAM each. Since we were training all model parameters, and for five folds for cross validation, these experiments were lengthy. However, given that full model fine-tuning was possible using two server-grade GPUs, model training is relatively accessible.

A.4 Textual Input/Prompt

We utilized a standard defintion + instruction format for the textual prompt for both datasets.

For AFLW, the input was:

> *Your task is to analyze the following image and characterize the head position from the following head positions list. head positions: [left, right, up, down, extremes, frontal]*
> *Definition of left: "The subject's head is turned towards the left side of the camera view."*
> *Definition of right: "The subject's head is turned towards the right side of the camera view."*
> *Definition of up: "The subject's head is tilted slightly back."*
> *Definition of down: "The subject's head is tilted slightly forward."*
> *Definition of extremes: "The head is turned or tilted greater than 90 degrees either to the left, right, up, or down of the camera view."*
> *Definition of frontal: "The subject's head is faced forward."*
> *Instructions: Only give a single head position label. Do not provide explanations, descriptions, or text outside of the requested format*

For CCDb-HG the input was:

> *Your task is to analyze the following video and characterize the present gesture from the following head gestures list. head gestures: [Nod, Shake, Tilt, Turn, Up/Down, None].*
> *Definition of Nod: "An up-down rotation along the pitch axis. It involves a slight, quick, or repetitive lowering and raising of the head."*
> *Definition of Shake: "A left-right horizontal rotation along the yaw axis. It involves a rapid and potentially repeated side-to-side motion, typically with small or moderate amplitude."*
> *Definition of Tilt: "A sideways rotation along the roll axis, involving a shift of the head in which one ear moves closer to the shoulder while the other ear moves*

away."
Definition of Turn: "A left or right rotation, involving the shifting of the head from its original position to another one facing a different direction."
Definition of Up/Down: "Similar to a turn, but along the pitch direction and usually involves a gaze shift in the same direction as the head."
Instructions: 1. Only give a single head gesture label. 2. Do not provide explanations, descriptions, or text outside of the requested format.

B Further Model Experiments and Evaluation

We additionally experimented with a VideoLLaMA 3 model trained on the full AFLW dataset and evaluated on CCDb-HG, and observe catastrophic forgetting (Luo et al., 2025) with a macro-F1 score of 0.000 compared to the base model performance of 0.108 on CCDb-HG. We also report here macro-F1 scores per label class in Tables 2 and 3.

C Dataset Details

AFLW AFLW2000-3D (Zhu et al., 2016) is a dataset comprised of 2,000 facial images from the Annotated Facial Landmarks in the Wild (AFLW) dataset (Koestinger et al., 2011). We refer to this dataset as AFLW in this work. We grouped images into six head pose categories based on yaw and pitch: frontal (1,471 images), right (217 images), left (232 images), up (21 images), down (32 images), and extreme (27 images), following Dao et al. (2024). These categories are fully defined by yaw and pitch, so we excluded roll, which holds limited discriminative values in 2D head pose analysis.

CCDb-HG The Cardiff Conversation Database-Head Gestures (CCDb-HG) (Vuillecard et al., 2024) contains short video clips extracted from video recordings of dyadic human conversations. Each clip features a single participant as the focus and captures all facial and nonverbal expressions and gestures. It contains 115 videos totaling 178k frames with 5K annotated events. The annotations are for six head gesture categories with the following imbalanced label distribution: *nod* (2,496 events), *shake* (848 events), *turn* (643 events), *tilt* (523 events), and *up/down* (248 events). Reproducing the pre-processing procedure of Vuillecard et al. (2024), we split videos into non-overlapping 31-frame sub-clips and discarded samples in the "waggle" category, yielding 17,514 gesture and non-gesture clips.

Experiment Type	Experiment Subtype	Train Dataset	Test Dataset	F1 score	F1 score - Down	F1 score - Extremes	F1 score - Frontal	F1 score - Left	F1 score - Right	F1 score - Up
Baseline (zero-shot)	–	–	AFLW	0.336175	0.23298	0	0.80004	0.39346	0.36172	0.17356
Single-task	Stage 2+3	AFLW	AFLW	0.61404	[HTML]FFFFFF0.36478	[HTML]FFFFFF0.2	0.97284	[HTML]FFFFFF0.89884	[HTML]FFFFFF0.9164	0.19

Table 2: VideoLLaMA 3 model experimental results: per-class F1 score performance for AFLW.

Experiment Type	Experiment Subtype	Train Dataset	Test Dataset	F1 score	F1 score - Nod	F1 score - None	F1 score - Shake	F1 score - Tilt	F1 score - Turn	F1 score - Up/Down
Baseline (zero-shot)	–	–	CCDb-HG	0.108	0.20925	0.178375	0.10155	0.0501	0.116275	0
Single-task	Stage 3+4	CCDb-HG	CCDb-HG	0.442	0.2847	0.8740	0.1730	0.2722	0.5587	0.3258

Table 3: VideoLLaMA 3 model experimental results: per-class F1 score performance for CCDb-HG.