VERIRL: GROUNDING VIDEO REASONING FOR DEPRESSION ASSESSMENT WITH VERIFIABLE FACIAL-DYNAMICS REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has shown strong potential for improving reasoning in multimodal large language models (MLLMs), yet extending it to medical video reasoning is challenging: task labels verify final predictions but not whether the behavioral evidence reported alongside them agrees with observable video evidence. We propose **VeriRL**, a post-training framework that grounds video reasoning through verifiable facial-dynamics rewards. Selected temporal facial Action Unit (AU) dynamics serve as measurable, clinically motivated behavioral references for auxiliary verification. To support progressive training, we construct two facial-evidence-aligned Chain-of-Thought corpora, **GroundFace-CoT** ($\sim$4K images) and **GroundVideo-CoT** (1,130 video clips). During RL, the dynamics reward measures agreement between structured model reports and precomputed OpenFace-derived references. We further introduce Dynamics-Grounded Sign-GRPO (DGS-GRPO), which uses a fixed-sign classification advantage to preserve correctness feedback under homogeneous rollouts, while a bounded dynamics term adjusts its magnitude without reversing the correctness-determined sign. Experiments on AVEC 2014 and our curated Vlog dataset show gains over the supervised baseline in binary recognition, improving Accuracy/Macro-F1 from $0.61/0.61$ to $0.63/0.63$ on Vlog and from $0.57/0.51$ to $0.60/0.55$ on AVEC. On held-out Vlog videos, dynamics agreement further increases from 0.64 to 0.72, with exact channel agreement improving from 0.50 to 0.60. These results support measurable behavioral signals as auxiliary verification sources for evidence-grounded medical video reasoning.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have demonstrated strong capabilities in visual understanding and complex reasoning (Lin et al., 2024; Zhang et al., 2024; Zhi et al., 2026). Beyond supervised instruction tuning, reinforcement learning with verifiable rewards (RLVR) has emerged as a promising paradigm for improving reasoning through computable feedback (Guo et al., 2025; Feng et al., 2026; Zheng et al., 2026). Its success in mathematical, coding, and video tasks highlights the importance of verifiable signals during post-training.

Extending this paradigm to medical video reasoning exposes a **verification gap**: outcome-level labels can verify the final prediction, but not whether the behavioral evidence reported by the model agrees with observable video evidence. A model may therefore produce a correct prediction while describing facial behavior inaccurately, and linguistic coherence alone does not establish agreement with visual observations. This motivates a central question: *can clinically motivated, measurable behavioral cues provide auxiliary verification beyond final-answer correctness?* Such cues could complement task supervision by explicitly comparing model-reported behavior with externally computed references.

Facial dynamics provide one source of measurable behavioral evidence. Prior depression-assessment research has investigated textual (De Choudhury et al., 2013), acoustic (Cummins et al., 2013), physiological (Acharya et al., 2018), and facial-behavior signals (Wang et al., 2018; Xu et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2024; Yoon et al., 2022). However, using these signals as predictive inputs does not ensure that generated reports agree with the underlying behavior.

Facial Action Units (AUs) provide interpretable descriptions of localized facial movements and can be automatically estimated using OpenFace (Baltrušaitis et al., 2018). Motivated by reported reductions in positive facial expressivity (Falkenberg et al., 2012; Mosley et al., 2025), we use AU6/AU12 for positive-expression dynamics and AU25/AU26 for orofacial movement, with the latter serving as candidate behavioral proxies for orofacial manifestations of psychomotor slowing. Their amplitude and temporal variation yield structured reference targets for verification. Importantly, these cues are clinically motivated behavioral references rather than diagnostic biomarkers or clinical ground truth.

Building on this idea, we introduce **VeriRL**, a post-training framework that grounds video reasoning for depression assessment using verifiable facial-dynamics rewards. We construct **GroundFace-CoT** ($\sim$4K images) for localized facial understanding and **GroundVideo-CoT** (1,130 video clips) for temporal video reasoning, and progressively adapt Qwen2.5-VL-7B-Instruct through image-level SFT, video-level SFT, and RL. During RL, the dynamics reward measures agreement between structured model reports and precomputed OpenFace-derived references.

Standard group-centered optimization can suppress classification feedback when rollout groups become uniformly correct or incorrect. We therefore introduce Dynamics-Grounded Sign-GRPO (DGS-GRPO), using a fixed-sign classification advantage to preserve correctness feedback while a bounded dynamics term refines its magnitude without reversing the correctness-determined sign.

Experiments on AVEC 2014 (Valstar et al., 2014) and our curated Vlog dataset show gains over supervised adaptation in binary recognition: Macro-F1 improves by 0.04 on AVEC, 0.02 on Vlog Full, and 0.04 on Vlog Clean. Four-class severity performance remains comparable under substantially sparser ordinal supervision. Together, these results support measurable facial dynamics as auxiliary verification signals for evidence-grounded video post-training.

Our contributions are:

- We formulate temporal facial AU dynamics as measurable auxiliary verification targets, complementing final-answer supervision with explicit behavioral-evidence consistency.
- We introduce DGS-GRPO, combining a fixed-sign classification advantage with a bounded dynamics adjustment to preserve correctness feedback under homogeneous rollouts while incorporating evidence-based refinement.
- We construct **GroundFace-CoT** and **GroundVideo-CoT** for progressive image-to-video grounding, and demonstrate gains in both binary recognition and held-out behavioral-evidence agreement over supervised adaptation.

2 Related Work

2.1 Reinforcement Learning for Multimodal Reasoning

Supervised instruction tuning supports video-language alignment (Lin et al., 2024; Zhang et al., 2024). VideoEspresso (Han et al., 2025) provides fine-grained video CoT annotations and core-frame selection, Video-R1 (Feng et al., 2026) introduces temporal-aware RL, and VITAL (Zhang et al., 2026a) combines visual tool use with difficulty-aware optimization for long-video reasoning. These approaches advance video reasoning and post-training, but outcome-level supervision does not directly evaluate whether reported behavioral evidence agrees with externally measured video evidence. VeriRL targets this complementary verification problem.

2.2 Verifiable Reward Learning for Reasoning Alignment

Rule-based verification has enabled reasoning-oriented post-training, as illustrated by DeepSeek-R1 (Guo et al., 2025), but constructing verifiable signals becomes harder for open-ended outputs (Yuan et al., 2026). Model-based or preference-oriented supervision does not inherently establish agreement with external evidence (Zheng et al., 2026), which is particularly relevant when medical explanations are plausible but unsupported (Jing et al., 2026). Recent work evaluates structured clinical decision paths against textual evidence (Pronesti et al., 2026). VeriRL instead uses ex-

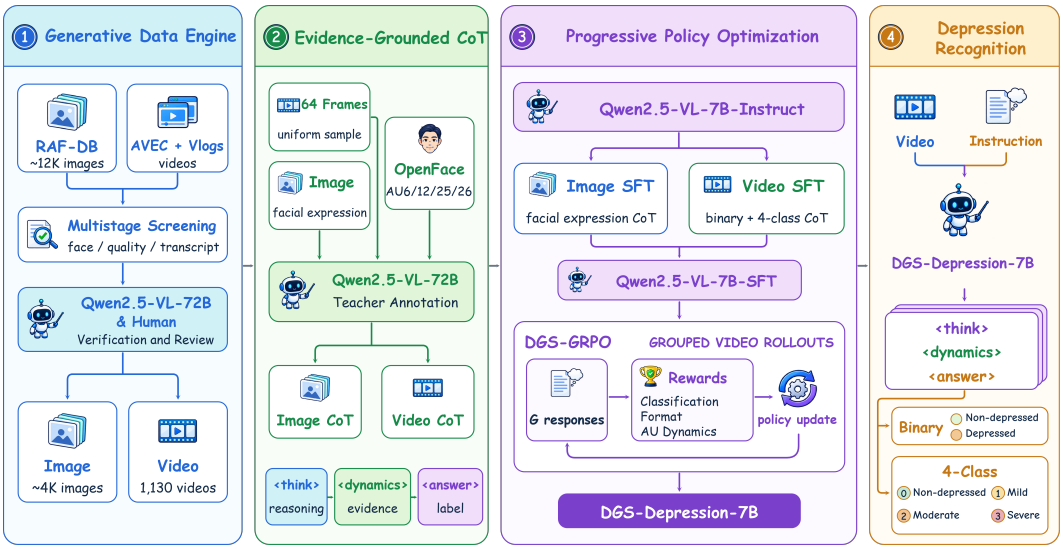


Figure 1: End-to-end pipeline of VeriRL. The framework consists of four stages: (1) a generative data engine yielding GroundFace-CoT (∼4K images) and GroundVideo-CoT (1,130 videos), (2) evidence-grounded CoT generation using OpenFace AU tracking, (3) progressive SFT and DGS-GRPO optimization, and (4) downstream depression assessment with the aligned DGS-Depression-7B model.

ternally computed facial dynamics to evaluate structured behavioral reports, providing AU-derived auxiliary rewards alongside task-label supervision.

2.3 Medical Video Reasoning and Depression Assessment

Depression assessment has explored textual, acoustic, physiological, and facial-behavior signals (De Choudhury et al., 2013; Cummins et al., 2013; Acharya et al., 2018; Wang et al., 2018). Video approaches model facial behavior across temporal scales (Xu et al., 2024); AVEC provides established benchmark settings (Valstar et al., 2013; 2014), while D-Vlog extends nonverbal analysis to naturalistic online videos (Yoon et al., 2022). MLlm-DR (Zhang et al., 2026b) further combines multimodal prediction with natural-language assessment rationales. Unlike approaches that primarily use facial signals as predictive inputs, VeriRL formulates selected temporal AU dynamics as auxiliary reward targets for structured behavioral-evidence consistency during MLLM post-training.

3 Method

We present **VeriRL**, a post-training framework for grounding video reasoning with verifiable facial-dynamics rewards. As illustrated in Figure 1, VeriRL constructs measurable behavioral references, progressively adapts the policy from image to video, and incorporates evidence consistency into reinforcement learning.

3.1 Problem Formulation and Design Principles

Let $\mathcal{X}$ denote the video input space and $\mathcal{Y}$ the task-label space. Given a video $x \in \mathcal{X}$, an MLLM policy π_θ generates a structured response $o \in \mathcal{O}$ with three sequential blocks: a reasoning rationale (<think>), a structured behavioral report (<dynamics>), and a final prediction (<answer>). Outcome supervision evaluates the final prediction, whereas VeriRL additionally evaluates whether the structured dynamics report agrees with externally computed behavioral references. Figure 5 shows an example. Our design follows two principles:

- **DP 1: Verifiable Evidence Grounding.** Behavioral evidence is evaluated against externally computed facial-dynamics references rather than linguistic plausibility alone.

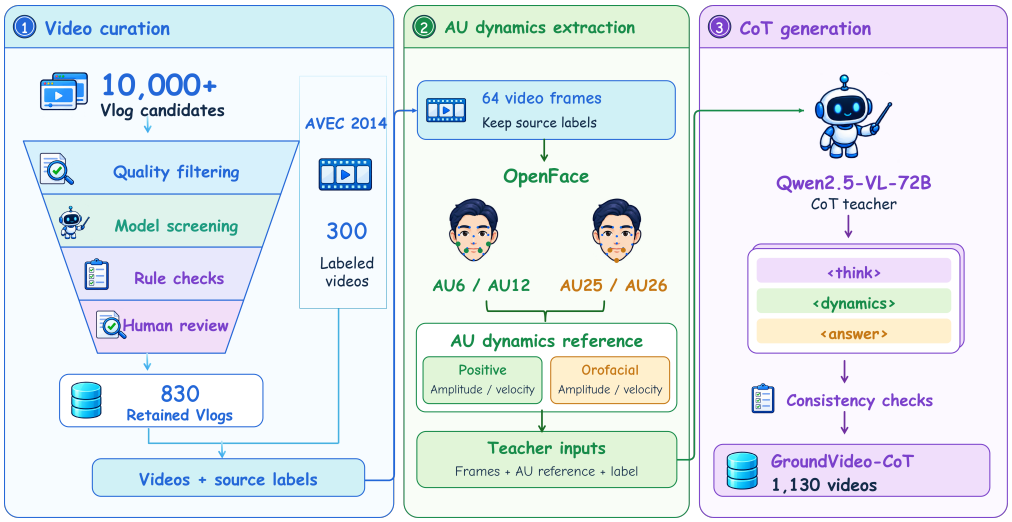


Figure 2: Construction of the facial-evidence-grounded GroundVideo-CoT dataset, including candidate acquisition, quality filtering, and SFT/RL data partitioning.

- **DP 2: Correctness-Prioritized Optimization.** Final-answer correctness determines the direction of the learning signal; evidence consistency may refine its magnitude but cannot reverse the sign determined by correctness.

3.2 Verifiable Facial Evidence Representation

To instantiate **DP 1**, we use facial Action Units (AUs) as measurable behavioral references. We construct **GroundFace-CoT** from RAF-DB (Li et al., 2017) for image-level facial grounding and **GroundVideo-CoT** from AVEC 2014 (Valstar et al., 2014) and curated Vlogs for temporal video reasoning. Figure 2 summarizes the construction pipeline; complete filtering and annotation details are provided in Appendix C. For video-level verification, OpenFace (Baltrušaitis et al., 2018) extracts temporal AU measurements. Prior findings on altered positive facial expressivity motivate AU6/AU12-based references (Falkenberg et al., 2012; Mosley et al., 2025). We additionally use AU25 (lips part) and AU26 (jaw drop) for orofacial movement as candidate behavioral proxies for orofacial manifestations of psychomotor slowing. These cues serve as behavioral references for auxiliary supervision rather than clinically validated diagnostic biomarkers. We summarize positive-expression (AU6/AU12) and orofacial (AU25/AU26) activity using amplitude and temporal-variation descriptors, yielding four reference channels. Each channel is standardized using training-set statistics and discretized into `low`, `mid`, or `high`; formal definitions are provided in Appendix C.6. GroundVideo-CoT comprises 1,130 videos; its training split contains 930 videos (200 AVEC with BDI-II-derived severity labels and 730 Vlogs with binary labels), from which a 400-video subset is used for RL. GroundFace-CoT contains approximately 4K valid image CoTs, with the class-balanced SFT subset detailed in Appendix C.

3.3 VeriReward: Task Correctness and Evidence Consistency

VeriReward separates two complementary signals: *Task Correctness* and *Evidence Consistency*. For a fixed input video x, let o_i denote its i-th rollout response. Let $\hat{e}_{i,k}$ denote the predicted level of dynamics channel k in o_i, and e_k the corresponding video-level reference. Let $\mathcal{C}_i$ denote the set of valid parsed channels. We define the dynamics agreement reward $R_{\text{dyn},i} \in [0,1]$ as

$$R_{\text{dyn},i} = \frac{\sum_{k\in\mathcal{C}_i} w_k \, g(\hat{e}_{i,k}, e_k)}{\sum_{k\in\mathcal{C}_i} w_k}, \qquad \mathcal{C}_i \neq \emptyset. \tag{1}$$

We use fixed channel weights $(0.35, 0.20, 0.30, 0.15)$ for positive-expression amplitude, positive-expression temporal variation, orofacial amplitude, and orofacial temporal variation, respectively.

These weights are fixed training hyperparameters and are not tuned on test data. The ordinal matching function assigns partial credit to adjacent levels:

$$g(\hat{e}_{i,k}, e_k) = \begin{cases} 1, & \hat{e}_{i,k} = e_k, \\ 0.5, & \hat{e}_{i,k} \text{ and } e_k \text{ are adjacent}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Let $c_i = \mathbf{1}[\hat{q}_i = q_i]$, with $c_i = 0$ for an unparseable answer, and let $v_i = \mathbf{1}[\mathcal{C}_i \neq \emptyset]$ indicate whether the dynamics report can be scored. We transform dynamics agreement into a centered adjustment:

$$\Delta_{\mathrm{dyn},i} = (2R_{\mathrm{dyn},i} - 1)v_i, \tag{3}$$

with $\Delta_{\mathrm{dyn},i} = 0$ when no valid channel is available. Thus, agreement above the fixed reference 0.5 yields a positive adjustment, whereas lower agreement yields a negative adjustment.

3.4 DGS-GRPO: Fixed-Reference Advantage Optimization

For each video x, the rollout policy samples $G = 8$ responses, $\{o_i\}_{i=1}^{G} \sim \pi_{\mathrm{old}}(\cdot \mid x)$. Standard GRPO constructs group-relative advantages:

$$A_i^{\mathrm{GRPO}} = \frac{R_i - \overline{R}}{\sigma_R + \epsilon}, \tag{4}$$

where R_i denotes the scalar reward assigned to response o_i. If the correctness component is homogeneous across a rollout group, meaning every response within the group is either fully correct or fully incorrect, group centering yields no relative classification signal. For our setting, however, correct and incorrect predictions should remain explicitly rewarded or penalized independently of other responses in the group. We therefore introduce **Dynamics-Grounded Sign-GRPO (DGS-GRPO)**, replacing the group-relative classification reference with a fixed-reference label term:

$$A_i^{\mathrm{DGS}} = \mathrm{clip}_{[-1,1]}\left[L_i + \lambda_{\mathrm{dyn}}\Delta_{\mathrm{dyn},i}\right], \tag{5}$$

where L_i is the task-specific label term and λ_{dyn} controls evidence refinement. Because L_i is evaluated against a fixed reference rather than the group mean, its classification signal remains nonzero even for homogeneous correctness groups. For binary depression recognition,

$$L_i^{\mathrm{bin}} = \alpha_{\mathrm{bin}}(2c_i - 1), \tag{6}$$

where $\alpha_{\mathrm{bin}} = 0.90$. With $\lambda_{\mathrm{dyn}} = 0.10$ and $\Delta_{\mathrm{dyn},i} \in [-1, 1]$, the evidence adjustment cannot reverse the correctness-determined sign: correct predictions receive advantages in $[0.8, 1]$, whereas incorrect or unparseable predictions receive advantages in $[-1, -0.8]$. For four-class severity prediction, let $d_i = |\hat{q}_i - q_i|$, where q_i is the severity category derived from the AVEC BDI-II score (Appendix C). We define

$$L_i^{\mathrm{4cls}} = \begin{cases} +\alpha_{\mathrm{exact}}, & \hat{q}_i = q_i, \\ -\alpha_{\mathrm{adj}}, & \hat{q}_i > 0,\ q_i > 0,\ d_i = 1, \\ -\alpha_{\mathrm{far}}, & \hat{q}_i > 0,\ q_i > 0,\ d_i \geq 2, \\ -\alpha_{\mathrm{cross}}, & \text{otherwise.} \end{cases} \tag{7}$$

This asymmetric construction assigns smaller penalties to within-depression ordinal errors while prioritizing the non-depressed/depressed boundary. Since $\lambda_{\mathrm{dyn}} = 0.02 < \min(\alpha_{\mathrm{exact}}, \alpha_{\mathrm{adj}}, \alpha_{\mathrm{far}}, \alpha_{\mathrm{cross}}) = 0.08$, the dynamics adjustment cannot reverse the sign of any four-class label term. Task-specific coefficients are reported in Section 4.

3.5 Progressive Alignment Pipeline

We initialize VeriRL from Qwen2.5-VL-7B-Instruct and perform three sequential stages:

- **Stage 1: Local Facial SFT.** We fine-tune on GroundFace-CoT to establish localized facial grounding using

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_t \log \pi_\theta(o_t \mid o_{<t}, x). \tag{8}$$

- **Stage 2: Video SFT.** We fine-tune on GroundVideo-CoT for structured temporal facial reasoning, yielding Qwen2.5-VL-7B-SFT.
- **Stage 3: VeriReward Alignment.** Starting from the SFT policy, we optimize DGS-GRPO on the 400-video RL subset using a clipped policy objective with KL regularization.

Let $h_{i,t} = (x_i, o_{i,<t})$ and define

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid h_{i,t})}{\pi_{\text{old}}(o_{i,t} \mid h_{i,t})}. \tag{9}$$

Writing $A_i = A_i^{\text{DGS}}$, the clipped surrogate is

$$u_{i,t} = \min\left\{\rho_{i,t}A_i,\ \text{clip}(\rho_{i,t}, 0.8, 1.3)A_i\right\}, \tag{10}$$

$$\widetilde{u}_{i,t} = \begin{cases} \max(u_{i,t}, 3A_i), & A_i < 0, \\ u_{i,t}, & A_i \geq 0. \end{cases} \tag{11}$$

The lower bound for negative advantages limits the update when the importance ratio becomes large, providing an asymmetric trust region. The RL objective is

$$\mathcal{J}_{\text{RL}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\widetilde{u}_{i,t} - \beta\mathcal{L}_{\text{KL}}, \tag{12}$$

where N is the number of responses, T_i the number of valid response tokens, and β controls KL regularization. The frozen SFT policy π_{ref} supplies the KL reference, whereas π_{old} supplies the importance ratio. For $q_{i,t} = \pi_{\text{ref}}(o_{i,t} \mid h_{i,t})/\pi_\theta(o_{i,t} \mid h_{i,t})$, $\mathcal{L}_{\text{KL}}$ averages $q_{i,t} - \log q_{i,t} - 1$ over tokens and responses, with numerical clipping. Task-specific optimization coefficients are reported in Section 4. The resulting aligned model is denoted **DGS-Depression-7B**.

4 Experiments

4.1 Experimental Setup

Model Initialization. Unless otherwise specified, all experiments use Qwen2.5-VL-7B-Instruct as the base model. We employ the same tokenizer, processor, and video preprocessing pipeline throughout training and evaluation. For each video, 64 frames are sampled along the valid-frame timeline and resized such that the longer side is 448 pixels.

Progressive Supervised Fine-Tuning. We first fine-tune on the 1,999-sample class-balanced subset of **GroundFace-CoT** to establish localized facial grounding, followed by **GroundVideo-CoT** for temporal video reasoning. Parameter-efficient fine-tuning uses LoRA with rank $r = 32$, scaling factor $\alpha = 64$, and dropout 0.05, while the visual projector remains trainable. Learning rates are 5×10^{-5} for LoRA parameters and 1.5×10^{-5} for the visual projector. Optimization uses AdamW with cosine learning-rate scheduling and a warmup ratio of 0.1.

Evidence-Quality Subsets. For Vlog analysis, *Clean* denotes videos with at least 64 valid OpenFace-tracked frames, whereas *Non-Clean* contains videos requiring repeated sampling from 16–63 valid frames. This partition is determined by preprocessing quality independently of model predictions.

VeriRL Alignment. Starting from the merged SFT checkpoint, we perform RL using DGS-GRPO on the designated 400-video training subset. During RL, only language-side LoRA parameters are optimized, while the visual encoder and projector remain frozen. For each input, the policy samples $G = 8$ candidate responses. For binary recognition, we use $\alpha_{\text{bin}} = 0.90$, $\lambda_{\text{dyn}} = 0.10$, and $\beta = 0.08$. For four-class severity prediction, $(\alpha_{\text{exact}}, \alpha_{\text{adj}}, \alpha_{\text{far}}, \alpha_{\text{cross}}) = (0.98, 0.08, 0.18, 0.98)$, $\lambda_{\text{dyn}} = 0.02$, and $\beta = 0.06$.

Evaluation Protocol. Candidate responses are sampled stochastically during RL, whereas greedy decoding is used for validation and testing. We report Accuracy and Macro-F1 for task performance. All methods use identical preprocessing, decoding, and answer-parsing procedures. Training is performed on two NVIDIA RTX PRO 6000 GPUs.

Table 1: Depression-assessment performance on the AVEC 2014 and Vlog test sets. Acc. denotes Accuracy and F1 denotes Macro-F1. Bold indicates the best value in each column, including ties.

	AVEC 2014				Vlog Binary					
	Four-class		Binary		Clean		Non-Clean		Full	
Model	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Gemini-2.5-Flash	0.17	0.15	0.50	0.35	0.54	0.47	0.52	0.47	0.53	0.47
Qwen2.5-VL-7B	0.20	0.08	0.50	0.33	0.50	0.42	0.54	0.46	0.52	0.43
Qwen3-VL-4B	0.24	0.23	0.51	0.45	0.54	0.48	**0.58**	0.56	0.56	0.52
Qwen3-VL-8B	0.26	0.19	0.54	0.48	0.48	0.44	**0.58**	0.56	0.53	0.50
LLaVA-OneVision-7B	0.50	0.17	0.50	0.33	0.54	0.44	0.52	0.40	0.53	0.42
VILA-1.5-8B	0.50	0.17	0.50	0.33	0.50	0.33	0.50	0.33	0.50	0.33
VideoLLaMA2	0.45	0.21	0.49	0.43	0.52	0.40	0.54	0.46	0.53	0.43
VideoLLaMA3	0.20	0.08	0.50	0.33	0.50	0.33	0.50	0.33	0.50	0.33
Qwen2.5-VL-7B-SFT	0.51	**0.28**	0.57	0.51	0.64	0.64	**0.58**	0.57	0.61	0.61
DGS-Depression-7B	**0.52**	**0.28**	**0.60**	**0.55**	**0.68**	**0.68**	**0.58**	**0.58**	**0.63**	**0.63**

4.2 Main Results

Table 1 compares VeriRL with general-purpose MLLMs and its supervised initialization on AVEC 2014 and Vlog.

Supervised Adaptation. Domain-specific SFT substantially improves over the base Qwen2.5-VL-7B model. AVEC four-class Accuracy/Macro-F1 increases from $0.20/0.08$ to $0.51/0.28$, while Vlog Full increases from $0.52/0.43$ to $0.61/0.61$. These results establish the progressively supervised model as a strong task-adapted baseline for evaluating the additional effect of RL alignment.

VeriRL Alignment. Building on SFT, VeriRL improves AVEC binary Accuracy/Macro-F1 from $0.57/0.51$ to $0.60/0.55$. On Vlog Full, performance increases from $0.61/0.61$ to $0.63/0.63$, with a larger improvement from $0.64/0.64$ to $0.68/0.68$ on the Clean subset. On Non-Clean videos, the change is smaller ($0.58/0.57$ to $0.58/0.58$). Together, the binary results show modest but consistent gains over the supervised initialization under the reported configuration.

Table 2: Component analysis of VeriRL on the Vlog binary test set.

Configuration	Accuracy	Macro-F1
SFT w/o GroundFace-CoT	0.57	0.56
SFT	0.61	0.61
GRPO w/o Dynamics	0.57	0.56
VeriRL	**0.63**	**0.63**

Sparse Ordinal Supervision. On AVEC four-class severity prediction, VeriRL remains comparable to SFT, with Accuracy/Macro-F1 changing from $0.51/0.28$ to $0.52/0.28$. Unlike the binary settings, only 200 training videos are distributed across four ordinal levels. The current experiments therefore do not establish a clear RL benefit for fine-grained severity prediction. Whether this limitation arises from data scale, reward design, or their interaction requires further investigation.

4.3 Component Analysis

We examine the contributions of progressive facial grounding and dynamics-based RL feedback on the Vlog binary benchmark.

Progressive Facial Grounding. Removing GroundFace-CoT reduces Accuracy/Macro-F1 from $0.61/0.61$ to $0.57/0.56$. This supports the usefulness of image-level facial supervision before video-level adaptation, although the ablation does not isolate which specific aspects of image-level supervision account for the gain.

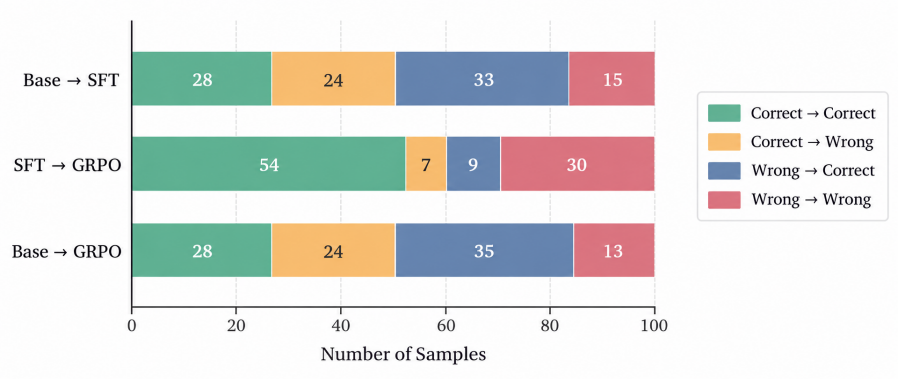


Figure 3: Prediction transitions across adaptation stages on the Vlog binary test set. Bars summarize samples that remain correct or incorrect and those whose correctness changes between the compared model stages.

Table 3: Held-out behavioral-evidence grounding on the Vlog binary test set. Dyn. Agree. denotes mean dynamics agreement; Exact denotes exact per-channel agreement; Correct-Ans. Agree. denotes dynamics agreement restricted to correctly classified samples.

Method	Dyn. Agree. ↑	Exact ↑	Correct-Ans. Agree. ↑	Acc. ↑
Qwen2.5-VL-7B-SFT	0.64	0.50	0.65	0.61
VeriRL	**0.72**	**0.60**	**0.74**	**0.63**

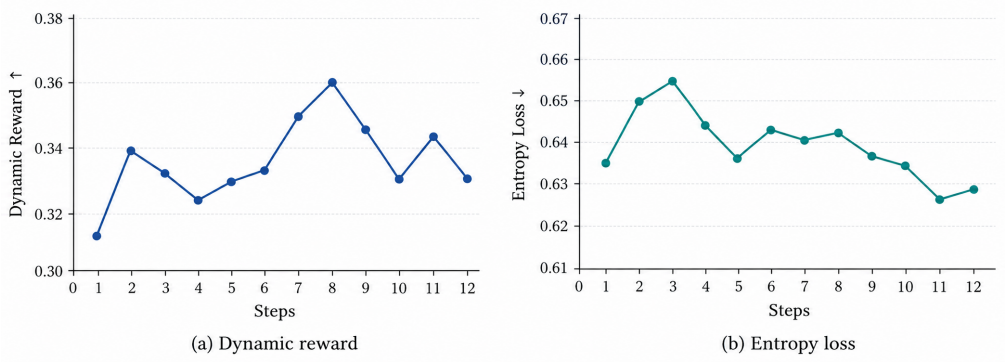
Dynamics-Based RL Feedback. Under the reported configuration, GRPO without dynamics reaches $0.57/0.56$, below its SFT initialization at $0.61/0.61$, whereas the complete VeriRL configuration reaches $0.63/0.63$. Thus, task-only RL does not improve over SFT in this experiment, while the complete configuration incorporating dynamics-based feedback yields the strongest binary performance. Because this ablation changes the dynamics component within the reported RL setup, it supports the utility of behavioral-evidence feedback but does not by itself establish the mechanism responsible for the degradation of task-only RL.

Figure 3 complements aggregate metrics by showing sample-level correctness transitions. The presence of both corrected and newly incorrect predictions illustrates that RL alignment changes individual decisions in both directions, even when aggregate performance improves.

4.4 Behavioral-Evidence Grounding

Task accuracy alone does not determine whether the structured behavioral evidence generated by a model agrees with the observed facial dynamics. We therefore directly evaluate predicted `<dynamics>` blocks against held-out OpenFace-derived references. These references are used only for evaluation and are not provided to the model during inference. We report three complementary metrics. **Dynamics Agreement** is the mean R_{dyn} defined in equation 1, which assigns partial credit to adjacent ordinal levels. **Exact Channel Accuracy** is the fraction of valid dynamics channels whose predicted level exactly matches its reference. **Correct-Answer Agreement** reports mean R_{dyn} restricted to correctly classified samples, testing whether evidence consistency improves even when final-answer correctness is held fixed.

As shown in Table 3, VeriRL improves held-out behavioral-evidence agreement beyond its gain in final-answer accuracy. Dynamics Agreement increases from 0.64 to 0.72, while Exact Channel Accuracy increases from 0.50 to 0.60. Importantly, Correct-Answer Agreement also increases from 0.65 to 0.74, showing that the improvement is not explained solely by the larger number of correct predictions: even among correctly classified samples, VeriRL produces structured dynamics reports that agree more closely with the external behavioral references. Together, these results provide direct held-out evidence that the dynamics-based RL objective improves the behavioral-evidence consistency it is designed to optimize, complementing the task-level gains in Table 1.



(a) Dynamic reward (b) Entropy loss

Figure 4: Training dynamics under DGS-GRPO, showing the dynamics-based reward and policy entropy during RL optimization.

4.5 Rationale Evaluation with AU References

Three LLM judges and human raters evaluate VeriRL-generated rationales using timestamped AU events and OpenFace-derived dynamics references from the corresponding videos. They assess AU-reference agreement and semantic quality using five criteria (Appendices C.10 and D.7). Table 4 reports mean overall scores of 4.38 – 4.84. This audit assesses whether behavioral claims in `<think>` are supported by video-derived references; it does not establish causal faithfulness or, without an SFT comparison, improvements attributable to RL.

4.6 Optimization Dynamics

Figure 4 shows the training-time dynamics reward and policy entropy during DGS-GRPO. The dynamics reward generally increases during early optimization before fluctuating within a relatively narrow range, showing that sampled structured reports receive progressively higher scores against the AU-derived references used by VeriReward. Policy entropy remains bounded during optimization and decreases toward the end of training. These curves characterize training behavior and should not be interpreted as held-out evidence of improved behavioral grounding.

Table 4: AU-reference-based rationale evaluation. Mean overall scores (1–5).

Scoring Method	Score (1–5)
GPT-5.4-mini	4.63
Gemini-2.5-Flash	4.84
DeepSeek-V4-Flash	4.38
Human Evaluation	4.60

5 Conclusion

We presented **VeriRL**, a post-training framework that incorporates measurable facial dynamics as auxiliary verification signals for video-based depression assessment. VeriRL separates final-answer correctness from structured behavioral-evidence consistency: OpenFace-derived temporal AU references provide a computable dynamics reward, while DGS-GRPO uses a fixed-reference classification advantage with bounded evidence-based refinement. GroundFace-CoT and GroundVideo-CoT further support progressive image-to-video grounding. Across AVEC 2014 and our curated Vlog benchmark, VeriRL yields modest but consistent gains over supervised adaptation in binary recognition and improves held-out agreement between structured dynamics reports and externally computed behavioral references, while four-class severity performance remains comparable under sparse ordinal supervision. Our results support measurable behavioral signals as a practical source of auxiliary verification for multimodal post-training, while further work is needed to establish broader task generalization and clinical validity.

Ethics statement

This study investigates video-based depression assessment using RAF-DB, AVEC, and vlog videos collected from Bilibili and YouTube. Facial AU dynamics serve as auxiliary behavioral references rather than diagnostic biomarkers, and model predictions and generated rationales should not be used as standalone clinical diagnoses. Facial videos and mental-health labels are sensitive; their use and any release of derived materials require careful consideration of privacy, consent, and source-data permissions. The present experiments do not establish clinical validity or fairness across demographic groups, and performance may vary with recording conditions and data sources. Any future clinical use would require independent validation and appropriate professional oversight, with safeguards against stigmatization and unauthorized mental-health profiling.

Reproducibility statement

Section 3 presents the structured response format, facial-dynamics reward, advantage construction, and policy-optimization objective. Section 4.1 reports the model initialization, training hyperparameters, and inference settings. Appendix C describes the data sources, annotation and filtering procedures, facial-dynamics representation, and construction of the CoT datasets. Appendix D provides prompt templates for data screening, CoT generation, and rationale evaluation using AU references. Code will be released upon publication. Access to source datasets and any redistribution of derived materials remain subject to the respective data-use permissions.

References

U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Hojjat Adeli, and D Puthankattil Subha. Automated eeg-based screening of depression using deep convolutional neural network. *Computer methods and programs in biomedicine*, 161:103–113, 2018.

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66. IEEE, 2018.

Nicholas Cummins, Julien Epps, Vidhyasaharan Sethu, Michael Breakspear, and Roland Goecke. Modeling spectral variability for the classification of depressed speech. In *Interspeech*, pp. 857–861, 2013.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pp. 128–137, 2013.

Irina Falkenberg, Nils Kohn, Regina Schoepker, and Ute Habel. Mood induction in depressive patients: A comparative multidimensional approach. *PLOS ONE*, 7:e30016, 2012.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26181–26191. IEEE, 2025.

Miao Jing, Mengting Jia, Junling Lin, Zhongxia Shen, Huan Gao, Mingkun Xu, and Shangyang Li. Beyond classification accuracy: Neural-medbench and the need for deeper reasoning benchmarks. In *International Conference on Learning Representations*, volume 2026, pp. 87125–87148, 2026.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2584–2593, 2017.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Philip E. Mosley, Johan N. van der Meer, Lachlan H. W. Hamilton, Jurgen Fripp, Stephen Parker, Jayson Jeganathan, Michael Breakspear, Richard Parker, Rebecca Holland, Brittany L. Mitchell, Enda Byrne, Ian B. Hickie, Sarah E. Medland, Nicholas G. Martin, and Luca Cocchi. Markers of positive affect and brain state synchrony discriminate melancholic from non-melancholic depression using naturalistic stimuli. *Molecular Psychiatry*, 30:848–860, 2025. doi: 10.1038/s41380-024-02699-y.

Massimiliano Pronesti, Anja Belz, and Yufang Hou. Beyond outcome verification: Verifiable process reward models for structured reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32187–32202, 2026.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 3–10, 2013.

Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, 2014.

Qingxiang Wang, Huanxin Yang, and Yanhong Yu. Facial expression video analysis for depression detection in chinese patients. *Journal of Visual Communication and Image Representation*, 57: 228–233, 2018.

Jiaqi Xu, Hatice Gunes, Keerthy Kusumam, Michel Valstar, and Siyang Song. Two-stage temporal modelling framework for video-based depression recognition using graph representation. *IEEE Transactions on Affective Computing*, 16:161–178, 2024.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12226–12234, 2022.

Zhonghang Yuan, Zhefan Wang, Fang Hu, Zihong Chen, Jinzhe Li, Gang Li, Jie Ying, Huanjun Kong, Songyang Zhang, and Nanqing Dong. Knowledge-to-verification: Exploring rlvr for llms in knowledge-intensive domains. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 40715–40749, 2026.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026a.

Wei Zhang, Juan Chen, En Zhu, Wenhong Cheng, YunPeng Li, Yuhan Li, and Yanbo J Wang. Mllm-dr: Towards explainable depression recognition with multimodal large language models. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22:1–23, 2026b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Xianda Zheng, Huan Gao, Meng-Fen Chiang, Michael J Witbrock, Kaiqi Zhao, and Shangyang Li. Evo-pi: Aligning medical reasoning via evolving principle-guided supervision. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 31835–31850, 2026.

Weihai Zhi, Jiayan Guo, and Shangyang Li. Medgr2: Breaking the data barrier for medical reasoning via generative reward learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28901–28909, 2026.

A AI USE STATEMENT

We used generative AI tools for candidate-data screening, synthetic CoT generation, and post-hoc rationale evaluation using AU references, as described in Appendices C and D. We also used generative AI tools for language editing of the manuscript and programming assistance. The authors take responsibility for the final manuscript, implementation, experimental results, and scientific claims, including content developed with AI assistance.

B LIMITATIONS

VeriRL has several limitations. First, four-class severity evaluation uses only 200 AVEC training videos across four ordinal levels, and the current experiments do not establish a clear RL benefit in this sparse setting. Second, VeriReward depends on OpenFace-derived AU measurements and may therefore degrade under occlusion, pose variation, or unreliable face tracking. Third, the selected AU dynamics are clinically motivated behavioral references rather than validated diagnostic biomarkers, and the current results do not establish clinical validity. Fourth, video-level dynamics verification is restricted to AU6, AU12, AU25, and AU26 and does not cover other facial actions investigated in depression research. Future work should evaluate broader AU sets and their combinations to assess their effects on behavioral-evidence consistency and task performance. Finally, evaluation is limited to depression assessment and does not yet establish whether the proposed verification strategy generalizes to other medical video tasks.

C DATA DESCRIPTION

C.1 OVERVIEW

Table 5 summarizes the training and test sample counts and the pooled binary-label distribution. Video counts refer to recordings, not necessarily distinct participants.

To support facial-behavior-based depression recognition and severity assessment, we construct two multimodal chain-of-thought corpora aligned with facial Action Unit (AU) dynamics: GroundFace-CoT and GroundVideo-CoT.

GroundFace-CoT is constructed from facial-expression images in RAF-DB. It is designed to teach the model fine-grained associations among localized facial muscle movements, AUs, and expression semantics. After image-quality filtering and CoT generation, we obtain approximately 4K valid image-level CoT samples. A class-balanced subset of 1,999 samples is selected for the first stage of supervised fine-tuning (SFT).

The video corpus contains 1,130 videos. The 300 AVEC 2014 interviews support four-class severity prediction, and 830 open-world Vlogs support binary recognition. AVEC labels are derived directly from the provided Beck Depression Inventory-II (BDI-II) scores. Vlog annotations are obtained through keyword retrieval, face-quality filtering, subtitle or automatic speech recognition (ASR), structured screening with the Qwen2.5-VL-72B-Instruct model, deterministic post-processing, and manual inspection.

The video splits are 200/100 train/test videos for AVEC 2014 and 730/100 train/test videos for Vlog. The complete video training set therefore contains 930 videos, and the test set contains 200 videos. A subset of 400 videos is further selected from the video training set for GRPO-based reinforcement learning. These 400 videos are part of the 930-video training set rather than an additional collection.

When the four AVEC categories are collapsed into a binary partition, the complete collection contains 565 non-depressed and 565 depressed videos. Mild, moderate, and severe AVEC samples are grouped into the depressed category for this aggregate statistic.

Table 5: Statistics of the datasets. Class distributions are reported over the complete dataset, including both training and test partitions.

		Dataset Split		Class Distribution	
Modality	Dataset	Train	Test	Non-dep.	Dep.
Facial Image	RAF-DB	1999	–	–	–
Video	AVEC	200	100	150	150
	Vlog	730	100	415	415

The construction process follows three principles. First, retrieval labels are treated only as weak source annotations and are never used directly as ground truth. Second, generated reasoning is constrained by externally measurable facial evidence: image CoT is grounded in static AU sets, whereas video CoT is grounded in four-channel temporal AU dynamics. Third, all statistics used for dynamics normalization and discretization are estimated from training data and remain fixed during validation and testing.

C.1.1 RAF-DB IMAGE DATA

GroundFace-CoT is constructed from the basic subset of RAF-DB, which provides seven facial-expression categories: anger, disgust, fear, happiness, neutral, sadness, and surprise. These images are used to learn localized facial movements, Action Unit (AU) evidence, and expression semantics. The seven expression categories are not mapped directly to depression labels; instead, they provide fine-grained facial-behavior supervision before video-level depression reasoning is introduced.

C.1.2 AVEC 2014 FOUR-CLASS DATA

AVEC 2014 is used for four-class depression-severity prediction. The benchmark consists of standardized interview videos and provides a Beck Depression Inventory-II (BDI-II) score for each sample. Unlike the open-world Vlog data, AVEC videos are not collected from YouTube or Bilibili, do not rely on keyword-derived weak labels, and are not relabeled using subtitles or vision-language-model outputs. Their severity labels are derived directly from the BDI-II scores provided by the benchmark.

Let b_i denote the BDI-II score of the i-th AVEC sample, and let $q_i \in \{0, 1, 2, 3\}$ denote its ordinal severity category:

$$q_i = \begin{cases} 0, & 0 \le b_i \le 13, \\ 1, & 14 \le b_i \le 19, \\ 2, & 20 \le b_i \le 28, \\ 3, & 29 \le b_i \le 63. \end{cases} \tag{13}$$

The four categories correspond to non-depressed, mild depression, moderate depression, and severe depression, respectively. They form an ordered label space, which is also reflected in the four-class DGS objective.

C.1.3 OPEN-WORLD VLOG BINARY DATA

The open-world Vlog dataset is used for binary depression recognition with non-depressed and depressed labels. Videos are collected primarily from YouTube and Bilibili. We prioritize personal vlogs and talking-head videos in which the primary subject appears on camera, speaks mainly about their own experience, and remains sufficiently visible throughout the clip. Candidate videos are generally restricted to a maximum duration of three minutes.

YouTube videos are retrieved with yt-dlp, together with available platform-provided or automatically generated subtitles. Bilibili videos are identified through platform search and downloaded using a compatible video-retrieval tool. The main YouTube collection period spans 2010–2026, whereas the Bilibili collection primarily contains videos published after 2012. To limit storage and computational costs, all videos are stored as MP4 files at resolutions no higher than 480p.

C.2 Open-World Vlog Collection

C.2.1 Search Keywords and Weak Labels

We use separate search-query groups for depression-related and ordinary daily-life content. Representative depression queries include "depression vlog," "living with depression," and "depressive episode vlog," among others. Representative daily-life queries include "daily vlog," "day in my life," and "morning routine." Semantically matched Chinese queries are also used for Bilibili retrieval.

Depression-related queries provide an initial depressed weak label, whereas daily-life queries provide an initial non-depressed weak label. These labels indicate only the retrieval source and are not treated as final annotations. All candidates therefore undergo content-based verification.

C.3 Open-World Vlog Screening

Candidate Vlogs undergo frame extraction, face detection, subtitle or speech transcription, structured vision-language screening, deterministic post-processing, and manual inspection. The subtitle-based procedure described in this subsection applies only to the open-world Vlog data; AVEC labels are determined directly from BDI-II scores.

Frame extraction and face detection. Candidate videos are initially sampled at 1 frame per second. Each extracted frame is resized such that its shorter side is 720 pixels. Face detection is performed using SCRFD from the InsightFace buffalo-l configuration. For a single-person frame, a face is considered sufficiently visible when its bounding-box area occupies at least 2% of the frame. For frames containing multiple faces, the primary face is required to occupy at least 5% of the frame.

Let N_{face} denote the number of frames containing a valid primary face and N_{all} denote the total number of sampled frames. The face-frame ratio is

$$\rho_{\text{face}} = \frac{N_{\text{face}}}{N_{\text{all}}}. \tag{14}$$

Only videos satisfying $\rho_{\text{face}} \geq 0.30$ are retained for content screening. This ratio measures the availability of a usable primary face, rather than how often the subject directly faces the camera. The threshold is applied uniformly across retrieval sources and candidate labels and is not used to assign class labels.

For each retained video, the timeline is divided into uniformly spaced temporal intervals. The median valid-face frame in each interval is selected whenever available; otherwise, the temporal center frame is used. This procedure produces up to 64 frames that cover the complete video while prioritizing face visibility.

Subtitles and ASR. For YouTube videos, available platform or automatically generated subtitles are used. When subtitles are unavailable, WhisperX generates a transcript. Each transcript uses timestamps in the form [mm:ss] text. Videos with fewer than 20 transcript characters are not passed to VLM screening.

Subtitles and ASR transcripts are used only for constructing Vlog annotations. They are not provided during video CoT generation, SFT, reinforcement learning, validation, or testing.

Vision-language-model screening. Videos that pass the initial quality gates are evaluated by a locally deployed Qwen2.5-VL-72B-Instruct model. Its inputs compriseup to 64 frames selected by the face-prioritized procedure above, the subtitle or ASR transcript, video identifier, duration, face-frame ratio, and keyword-derived weak label. It predicts depression, non-depression, or exclusion. The keyword-derived label is treated only as weak metadata.

The output is constrained by a strict JSON Schema. The main inference configuration uses BF16 precision, a maximum context length of 65,536 tokens, at most 64 visual inputs, temperature 0.1, a maximum output length of 2,000 tokens, a 120-second timeout, and at most two retries.

Criteria for positive depression samples. A video is labeled as depression-related only when all of the following conditions are satisfied: (1) the subject shown in the video is the person whose experience is being described; (2) the subject has a sufficiently visible face; (3) the content concerns a

current or recent depression-related state; (4) the transcript contains direct supporting evidence; and (5) at least one of the following evidence types is present: current depression-related treatment, persistent depression-related symptoms with a temporal anchor, current or recent self-harm or suicidal ideation, or a currently relevant depression-related diagnosis.

The mere occurrence of words such as "depression" or "depressed" is insufficient for a positive label. A video that discusses a past depressive episode and explicitly states that the speaker has recovered is not labeled as depressed. Conversely, a report of recent self-harm or suicidal ideation may support a positive annotation when the speaker explicitly indicates that the distress remains current.

Exclusion categories. Videos are excluded under the following conditions:

- No face: the primary face is absent, insufficiently visible, or rarely directed toward the camera;
- Third person: the video primarily discusses another person, a news event, an interview subject, or a third-party story;
- Recovered: the speaker describes only a resolved depressive episode or primarily provides recovery advice;
- Temporary negative affect: the video expresses temporary sadness, fatigue, frustration, heartbreak, or ordinary negative affect without persistent or clinically relevant evidence;
- Irrelevant content: the audiovisual content is unrelated to the task;
- Low quality: the video, transcript, or available evidence is insufficient for reliable annotation;
- Life-crisis distress: the speaker expresses severe distress related to a specific life event but provides no evidence of persistent depressive symptoms, clinical course, or diagnosis.

Face-visibility verification. The screening model additionally determines whether the same primary speaker appears throughout the video, whether the face is severely occluded, whether the subject is an avatar or non-real face, and the proportion of time for which the subject faces the camera. The face-to-camera ratio is divided into three levels: majority to camera (at least 70%), partially to camera (30%–70%), and rarely to camera (below 30%). Videos in the rarely-to-camera category are excluded because the available facial evidence is insufficient.

Videos with a face-to-camera ratio below 30% are excluded for insufficient facial evidence. This quality criterion is applied uniformly to all candidates, independently of their retrieval labels or eventual annotations. Videos that pass the common quality checks are then annotated using the content-evidence criteria and exclusion rules. Membership in the partially-to-camera category does not determine the class label.

Deterministic post-processing. Structured model predictions are further checked using deterministic rules. Severely occluded or virtual subjects are excluded for insufficient facial evidence. Videos with inconsistent speakers, news content, tutorials, or interview-style third-person content are excluded as third-person cases. Videos with a face-to-camera ratio below 30% are excluded regardless of their candidate labels. This quality filter determines eligibility only; it does not assign a depression or non-depression label.Positive samples without sufficiently detailed transcript evidence, non-excluded predictions with confidence below 0.5, and unparseable or internally inconsistent outputs are conservatively excluded. The remaining samples are manually inspected, and only depression and non-depression examples are retained.

Final Vlog dataset. We checked the Vlog training and test partitions for duplicate video content and participant overlap. No exact or near-duplicate videos, clips from the same original recording, or identified participants were shared across the two partitions.

The final Vlog dataset contains 830 videos: 730 training videos and 100 test videos, with 415 depressed and 415 non-depressed videos. The training set is used for video-level SFT and subsequent reinforcement learning. The test set is used only for final evaluation.

C.4 RAF-DB Construction and Filtering

Image preprocessing. The aligned RAF-DB images have an original resolution of 100×100. Two input versions are generated: images are padded to 140×140 for OpenFace and resized to 224×224 RGB for the teacher model.

AU extraction. OpenFace FaceLandmarkImg is used to extract image-level facial AUs. For the i-th image, the reference AU set is

$$G_i = \{a \in \mathcal{V}_{\mathrm{AU}} \mid \mathrm{AU}_c(i, a) = 1\}, \tag{15}$$

where

$$\mathcal{V}_{\mathrm{AU}} = \begin{Bmatrix} \mathrm{AU01}, & \mathrm{AU02}, & \mathrm{AU04}, & \mathrm{AU05}, \\ \mathrm{AU06}, & \mathrm{AU07}, & \mathrm{AU09}, & \mathrm{AU10}, \\ \mathrm{AU12}, & \mathrm{AU15}, & \mathrm{AU17}, & \mathrm{AU20}, \\ \mathrm{AU23}, & \mathrm{AU25}, & \mathrm{AU26} & \end{Bmatrix}. \tag{16}$$

When more than five AUs are active, the continuous intensity values AU_r are used to rank the active AUs, and the five strongest AUs are retained.

The reference AU set is used as teacher-side evidence, as the target content of the AU block, and as a deterministic constraint for detecting unsupported AU mentions.

Image-quality filtering. An image is retained only when OpenFace detection succeeds, the OpenFace confidence is at least 0.85, the Laplacian variance of the 224×224 image is at least 15, exactly one usable face is present, and the reference AU set is non-empty for non-neutral expressions.

A total of 7,331 RAF-DB training images pass these filters. Approximately 4,000 images are selected for CoT distillation, yielding 3,970 image CoT samples that pass online generation checks, 3,962 samples after format normalization, and 1,999 class-balanced samples for Stage-1 SFT. Class balancing uses random seed 3407. The final subset contains 157 fear samples and 307 samples for each of the other six expression categories.

C.5 Unified Video Preprocessing

Valid-frame selection and sampling. AVEC and Vlog videos are processed using the same visual pipeline after their labels have been determined. OpenFace is used for frame-level face tracking and AU extraction. A frame is considered valid when OpenFace detection succeeds, face-tracking confidence is at least 0.90, and the primary face is usable without unresolved multi-face ambiguity.

If at least 64 valid frames are available, 64 frames are uniformly sampled along the valid-frame timeline. If 16–63 valid frames are available, repeated sampling is used to obtain 64 inputs, and the sample is marked accordingly. Videos with fewer than 16 valid frames are excluded from facial-dynamics construction. Each selected frame is resized such that its longer side is 448 pixels. The same preprocessing pipeline is used during training, validation, and testing.

OpenFace AU signals. The temporal representation focuses on AU6 (cheek raiser), AU12 (lip corner puller), AU25 (lips part), and AU26 (jaw drop). AU6 and AU12 represent positive-expression activity, whereas AU25 and AU26 represent orofacial and jaw movement. For each AU group, frame-level intensities are converted into a channel-level temporal sequence using a fixed aggregation procedure shared across all data splits.

C.6 Four-Channel Facial Dynamics

Each video is represented by four dynamics channels: the amplitude and velocity of positive expression and orofacial movement. The positive-expression channels are derived from AU6 and AU12, whereas the orofacial channels are derived from AU25 and AU26.

Let u_t denote the channel-level AU intensity at the original video timestamp τ_t, measured in seconds, with $T \geq 2$ and $\tau_{t+1} > \tau_t$. We use the amplitude descriptor

$$a = \mathrm{Percentile}_{95}(u_t). \tag{17}$$

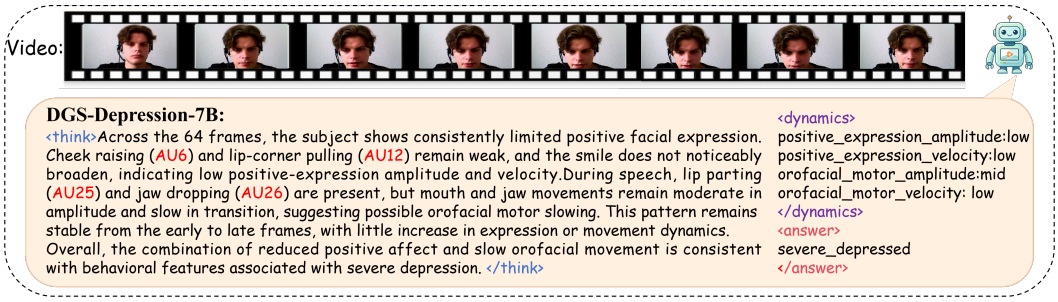


Figure 5: Example of structured video reasoning produced by VeriRL, including the behavioral report used for evidence-consistency evaluation.

The mean absolute temporal rate of change is

$$v = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{|u_{t+1} - u_t|}{\tau_{t+1} - \tau_t}. \tag{18}$$

Here, T counts valid observations at distinct original timestamps; repeated frames introduced solely to pad the model input are not additional observations. The quantity measures AU-intensity change per second, rather than physical facial displacement speed. These statistics yield the four-dimensional reference

$$E_{\text{ref}} = [e^1, e^2, e^3, e^4], \tag{19}$$

corresponding to the amplitude and velocity of positive-expression and orofacial movement.

For the i-th video and the k-th channel, the raw value e_i^k is standardized as

$$z_i^k = \frac{e_i^k - \mu_k}{\sigma_k}, \tag{20}$$

where μ_k and σ_k are estimated from the training data. The standardized values are discretized into three ordinal levels: low, mid, and high. These levels define the reference dynamics block, the evidence-matching target in VeriReward, and the privileged dynamics card used during teacher generation. All statistics and discretization boundaries remain fixed during validation and testing.

C.7 Image Chain-of-Thought Construction

GroundFace-CoT is generated with a locally deployed Qwen2.5-VL-72B-Instruct teacher model. Responses are generated in English with a maximum length of 1,024 tokens. Each response follows the structure

```
<think>Facial-region and muscle-movement
reasoning.</think>
<au>Reference AU list.</au>
<answer>Expression category.</answer>
```

The teacher receives the image, the reference AU set G_i, and the current candidate-emotion set, but not the ground-truth expression label. A response is accepted only when the predicted expression matches the RAF-DB label, the AU block exactly matches G_i, all AU mentions in the reasoning block belong to G_i, and all three output blocks are present and parseable.

If the predicted expression is incorrect, that category is removed from the candidate set and the teacher is prompted to reconsider the image. Each image permits at most six label-elimination rounds and two outer retries.

The teacher is instructed to describe the overall facial impression first, identify visibly active facial regions, describe muscular movements before naming their AUs, combine regional evidence, and infer the expression only at the end. The reference AU set is privileged teacher-side information and is removed from the student SFT prompt.

Table 6: Summary of the data used to construct GroundFace-CoT and GroundVideo-CoT. GroundVideo-CoT contains 1,130 videos, including 930 training videos and 200 test videos. A subset of 400 training videos is used for GRPO.

Modality	Dataset	Total	Train	Test	Task
Image	RAF-DB	∼4K CoT	1,999	–	Expression and AU learning
Video	AVEC 2014	300	200	100	Four-class severity prediction
Video	Vlog	830	730	100	Binary depression recognition

C.8 Video Chain-of-Thought Construction

AVEC and Vlog use the same video CoT generation framework. The Qwen2.5-VL-72B-Instruct teacher receives 64 video frames, the four-channel reference dynamics card E_{ref}, the task label, and the task-specific instruction, and generates English responses with a maximum length of 900 tokens. Subtitles, ASR transcripts, titles, and retrieval keywords are not provided. Using the video and the dynamics card as auxiliary behavioral guidance, the teacher reasons about positive-expression and orofacial movements and generates a complete response with `<think>`, `<dynamics>`, and `<answer>` blocks. The dynamics block expresses the teacher-generated account of amplitude and temporal variation as four categorical fields; it is produced through video-based reasoning conditioned on the supplied reference. The Vlog answer space is binary, whereas AVEC uses the four benchmark severity categories. Generated responses are retained only when all blocks are present and parseable, the dynamics fields match E_{ref}, the answer matches the task label, directional statements in the rationale agree with the reference, and the rationale ends with the required conclusion sentence. These checks enforce structural and reference consistency; the dynamics card and task label remain teacher-side supervision.

The teacher generates the complete response, including the reasoning, dynamics, and answer blocks:

```
<think>Temporal reasoning based on
positive-expression and orofacial movements.</think>
<dynamics>
positive_expression_amplitude: low|mid|high,
positive_expression_velocity: low|mid|high,
orofacial_motor_amplitude: low|mid|high,
orofacial_motor_velocity: low|mid|high
</dynamics>
<answer>Task label.</answer>
```

C.9 Supervised and Reinforcement-Learning Data

Stage-1 SFT uses the 1,999-sample class-balanced subset of GroundFace-CoT. Stage-2 video SFT uses 200 AVEC training videos and 730 Vlog training videos, for a total of 930 training videos. AVEC and Vlog use the same structured response format and dynamics representation but different label spaces.

A subset of 400 videos is selected from the video training set for reinforcement learning. This subset is optimized using classification correctness and facial-dynamics consistency. Output formatting is monitored separately and has zero reward weight in the reported run.

C.10 Rationale Evaluation with AU References

Evaluation inputs. We evaluate generated rationales using GPT-5.4-mini, Gemini-2.5-Flash, DeepSeek-V4-Flash, and human raters. All evaluators use the same reference-input protocol and rubric. Each evaluator receives the complete generated response, timestamped AU events, and the four-channel OpenFace-derived dynamics reference for the corresponding video. The generated answer is available for checking internal consistency, whereas the ground-truth task label and model identity are withheld. The original video is not supplied to the evaluators.

Reference information. AU-event records specify the covered AUs, event intervals in original video time, and available activation or change descriptors. Tracking-quality information is included when available. The global dynamics reference uses the four ordinal channels defined in Appendix C.6. Reference extraction does not use the generated response or the task label. Global dynamics categories are not treated as evidence for an exact event time or temporal ordering.

Claim-level assessment. Evaluators identify observable facial-behavior claims in `<think>` and classify each as `supported`, `contradicted`, or `not_verifiable`. Each judgment identifies the relevant AU event, time interval, or reference channel. A claim is contradicted only when it clearly conflicts with reliable supplied evidence. Missing reference coverage or insufficient tracking quality is recorded as not verifiable, rather than as a factual error. Clinical interpretations and the final task label are not treated as directly verifiable facial observations.

Scoring. The five dimensions are logical coherence, internal consistency, AU-reference agreement, factual and clinical restraint, and relevance/specificity, each scored from 1 to 5. AU-reference agreement is assessed using observable claims that can be evaluated from the supplied references. If no observable claim is assessable, AU-reference agreement and the overall score are recorded as not assessable.

For assessable responses, the overall score is the arithmetic mean of the five dimensions after applying prespecified caps. The score is capped at 2 for a central behavioral claim clearly contradicted by reliable references, a serious contradiction among the generated blocks, an unjustified definitive clinical diagnosis, or a lack of meaningful reasoning. Predicting a task label does not by itself constitute a definitive clinical diagnosis. The complete rubric is provided in Appendix D.7.

Aggregation and use. For each evaluator, scores are averaged over assessable responses; responses with no assessable claims are counted separately. The mean overall scores are 4.63, 4.84, 4.38, and 4.60 for GPT-5.4-mini, Gemini-2.5-Flash, DeepSeek-V4-Flash, and human evaluation, respectively. This audit is used only for post-hoc evaluation and does not serve as a training reward or a hard filter for training samples.

C.11 Illustrative Model Output

Figure 5 illustrates the three-block response generated by the trained VeriRL model on a video example. The example presents the output format rather than an aggregate evaluation of behavioral grounding.

D Prompt Templates

D.1 Vlog Content-Screening Prompt

D.1.1 System Prompt

You are a strict annotator for depression-related short videos and follow the data-screening principles of D-Vlog.

Your task is to classify an open-domain personal Vlog of no more than three minutes into exactly one of the following categories:

- depression;
- non_depression;
- exclude.

[Input Description]

You will receive up to 64 key frames uniformly sampled along the video timeline; the complete platform subtitle or automatic speech recognition transcript; video_id; duration_sec; keyword_label; and face_frame_ratio.

The keyword_label only indicates the search-query group through which the video was retrieved. It is a weak source label and must not be used directly as the final classification label.

[Speaker and Face Verification]

Complete the following fields in observations.face_check:

1. same_speaker_throughout: Determine whether the primary subject remains the same speaker throughout the video.
2. occluded: Determine whether the subject's facial features are severely obscured by a mask, sunglasses, hair, poor lighting, or another form of occlusion.
3. is_avatar: Determine whether the subject is an animated character, virtual avatar, VTuber, or another non-real human face.
4. speaker_to_camera_ratio: majority_to_camera means that the subject faces and speaks toward the camera for at least 70% of the video; partially_to_camera means 30%–70%; and rarely_to_camera means less than 30%.

If `speaker_to_camera_ratio` is `rarely_to_camera`, output `exclude` and set `exclude_reason` to `no_face`. Apply this rule equally to all candidates.

If the value is `partially_to_camera` or `majority_to_camera`, continue with the content-evidence assessment and the remaining exclusion checks. Do not assign `depression` or `non_depression` from the face-to-camera category.

[Core Classification Principles]

Principle 1: Mentioning depression does not constitute a positive depression example. The appearance of words such as "depression," "depressed," or semantically similar expressions does not by itself justify a depression label. Determine whether the speaker is describing a depression-related state that is current at the time of recording or occurred recently and remains relevant.

Principle 2: Tense and the primary purpose of the video are critical. Output exclude and set exclude_reason to recovered if any of the following conditions holds:

- the speaker only describes a past depressive experience and explicitly states that the condition has improved;
- the primary purpose of the video is to share recovery strategies, successful treatment experiences, or advice for other people;
- the transcript contains clear recovery signals, such as "I no longer...," "I am much better now...," "how I overcame depression," or semantically equivalent phrases.

However, a recent report of self-harm or suicidal ideation may still serve as current evidence when the speaker explicitly states that they remain in distress.

Principle 3: The evidence must be directly related to depression. Valid evidence must concern at least one of the following:

- depression or a related diagnosis;
- persistent depression-related symptoms;
- current or recent self-harm or suicidal ideation;
- depression-related treatment.

The following must not be treated as evidence of depression:

- medication withdrawal unrelated to depression;
- chronic pain or insomnia not explicitly attributed to depression;
- treatment for unrelated physical or psychiatric conditions;
- lyrics, films, scripted performances, voice acting, or fictional stories;
- purely spiritual or alternative-medicine discussions;
- metaphorical, humorous, or challenge-related mentions of suicide.

[Exclusion Categories]

Check all exclusion conditions before considering a depression label. If any condition is met, output exclude.

1. no_face: No valid primary face is visible, or the subject faces the camera for less than 30% of the video.
2. third_person: The video primarily discusses another person, news, educational content, an interview subject, a celebrity, or a third-person story.
3. recovered: The video only describes a past depressive period from which the speaker has recovered, or primarily presents recovery advice.
4. bad_mood_only: The speaker only describes fatigue, frustration, heartbreak, a temporary low mood, or general emotional distress without a duration anchor, clinically relevant terms, or self-harm evidence.
5. irrelevant_content: The visual or audio content is unrelated to the task, such as a music video, lyrics, scripted performance, cover song, landscape footage, or recitation.
6. low_quality: The video is blurred, silent, nearly black, text-only, has an insufficient transcript, or lacks adequate evidence.
7. life_crisis_distress: The speaker expresses intense distress caused by a specific event, such as unemployment, family conflict, legal problems, or bereavement, but provides no evidence of a depression diagnosis, persistent symptoms, or clinical course.

[Requirements for a Depression Label]

A video may be labeled depression only when both Condition A and Condition B are satisfied.

Condition A: The speaker is the person shown in the video and has a sufficiently clear and visible face.The same face-visibility requirements apply to all candidate labels. Failure to meet these requirements leads to `exclude`, not to `non_depression`.

Condition B: At the time of recording or in the recent period, the speaker explicitly reports at least one of the following:

1. Depression-related treatment. The evidence should include both a depression-related term and a treatment-related term, such as depression, MDD, doctor, medication, psychotherapy, CBT, psychiatrist, SSRI, or antidepressant.
2. Persistent depression-related symptoms with a duration anchor. The evidence should include a symptom term, such as insomnia, hypersomnia, lack of motivation, anhedonia, hopelessness, emptiness, worthlessness, or persistent low mood, together with a duration expression, such as weeks, months, years, "for the past...," chronic, long-term, or a semantically equivalent phrase.
3. Current or recent self-harm or suicidal ideation.
4. A currently relevant depression-related diagnosis, such as "diagnosed with depression," MDD, bipolar disorder, or a semantically equivalent explicit statement.

[Requirements for a non_depression Label]

The following may be classified as non_depression:

- ordinary daily-life Vlogs;
- study, work, travel, food, cooking, or gaming videos;
- temporary and ordinary emotional fluctuations;
- videos retrieved with depression-related keywords that do not satisfy the strict depression criteria;
- videos with a valid visible speaker that do not satisfy any exclusion criterion.

[Evidence Requirements]

The evidence_quote must be copied directly from the transcript and include a timestamp. For a depression sample, evidence_quote must contain an explicit depression-related or clinically relevant term, provide enough information to support the decision, and directly describe the speaker's current or recent state.

[Required Execution Order]

1. Complete observations.
2. Complete the face and speaker verification.
3. Extract evidence_quote.
4. Check all exclude conditions first.
5. Check Condition A and Condition B for depression.
6. Assign non_depression to the remaining valid samples.
7. If the evidence is insufficient or the fields are contradictory, output exclude and set exclude_reason to low_quality.

Return only a JSON object that follows the specified JSON Schema. Do not output any additional text.

D.1.2 USER INPUT TEMPLATE

```
video_id: {video_id}
duration_sec: {duration_sec}
keyword_label: {keyword_label}
face_frame_ratio: {face_frame_ratio}
n_frames_provided: {n_frames}
frame_sampling: uniform_time

Transcript: {transcript}

The following inputs are uniformly sampled frames from
the video: {image_inputs}
```

D.2 IMAGE COT GENERATION PROMPT

D.2.1 TEACHER PROMPT

You are an expert in the Facial Action Coding System (FACS) and facial-expression analysis. You are given a face image and the AUs detected by the system. Analyze the face like a human expert: first observe the overall appearance, then describe visible muscular changes, name the corresponding AUs, and finally infer the facial expression.

Candidate emotions: {candidate_emotions}.

Allowed AU vocabulary: AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU15, AU17, AU20, AU23, AU25, and AU26.

Face image: <image>. System-detected AUs: {detected_aus}.

[Reasoning Requirements]

1. Begin with the overall facial impression. Do not use AU codes in the overall-impression sentence.
2. Focus only on facial regions that exhibit visible muscular movement. Candidate regions include the brows, eyes, eyelids, nose, cheeks, upper lip, lips, chin, and jaw.
3. For each attended region, first describe the visible muscular movement or deformation, and then name the corresponding AU.
4. Every AU token appearing inside <think> must belong to {detected_aus}.
5. If a visible movement may correspond to an AU not contained in {detected_aus}, describe only the movement and do not write the corresponding AU code.
6. Report only AU presence. Do not report numerical values or AU intensities.
7. Use meaningful AU combinations and the absence of signature AUs to distinguish the selected expression from its most confusable alternative.
8. Do not infer the expression before completing the facial-region analysis.
9. The final answer must be exactly one expression from the candidate list.

[Output Format]

```
<think>
1. Overall impression: ...
2. Region-by-region analysis: ...
3. AU combination and category discrimination: ...
4. Conclusion: ...
</think>
<au>AU..., AU...</au>
<answer>emotion</answer>
```

Inside <au>, list all and only the system-detected AUs. If no AU is detected, output <au></au>. Do not add any text outside the three required blocks.

D.2.2 STUDENT PROMPT

The student prompt retains the face image, the seven candidate expressions, the closed AU vocabulary, the facial-region analysis requirements, and the three-block output format. It removes the system-detected AU set, the ground-truth expression label, and label-elimination feedback. The candidate expressions are anger, disgust, fear, happiness, neutral, sadness, and surprise.

D.3 VIDEO COT GENERATION PROMPT

D.3.1 SHARED SYSTEM PROMPT

You are an expert in facial-behavior analysis. Write in English and return exactly three blocks in order: <think>...</think>, <dynamics>...</dynamics>, and <answer>...</answer>. Use the supplied dynamics card and task label as references for generating a consistent response.

D.3.2 TEACHER PROMPT

You are analyzing a short video clip containing 64 uniformly sampled frames.

The objective facial-motion measurements for this subject are {dynamics_card}. The ground-truth task label is {ground_truth_label}. The task is {task_question}.

[Reasoning Requirements]

Observe the 64 frames and qualitatively describe the subject's facial movements. Discuss both the magnitude of the facial movements and the rate at which the facial movements change over time.

When mentioning a facial movement, include the corresponding Action Unit in parentheses, for example: cheek raise (AU6), lip-corner pull (AU12), lips part (AU25), and jaw drop (AU26).

Use restrained language to relate the observations to relevant behavioral constructs. Possible relationships include reduced smiling being associated with reduced positive affect, limited facial-movement amplitude being associated with blunted affect, and slowly changing facial movements being associated with psychomotor slowing.

[Hard Constraints]

1. Do not use the tier words `low`, `mid`, or `high` inside the `<think>` block. Use qualitative descriptions instead. Use these tier words only as categorical field values inside the `<dynamics>` block.
2. Do not report numerical AU measurements.
3. The direction of the reasoning must agree with the supplied facial-dynamics reference.
4. Do not introduce unsupported facial movements or Action Units.
5. Do not claim that facial behavior alone constitutes a definitive clinical diagnosis.

6. End the <think> block with the following exact sentence: "{conclusion sentence}."
7. Return exactly one <think> block, one <dynamics> block, and one <answer> block, in that order.
8. Inside <dynamics>, report all four predefined dynamics fields using the existing output schema. Each field must contain one of low, mid, or high, matching the supplied dynamics card. Inside <answer>, output only the supplied task label. Do not include any text outside the three required blocks.

D.4 VLOG BINARY VIDEO COT PROMPT

D.4.1 TASK PROMPT

Decide whether the subject exhibits depression-related facial behavior. The candidate labels are non_depressed and depressed.

D.4.2 REQUIRED CONCLUSION SENTENCES

For a depressed sample, the reasoning must end with "Conclusion: depression-related facial behavior is present." For a non-depressed sample, the reasoning must end with "Conclusion: depression-related facial behavior is not present."

D.5 AVEC FOUR-CLASS VIDEO COT PROMPT

D.5.1 TASK PROMPT

Determine the subject's depression-severity level from the observed temporal facial behavior. The candidate labels, ordered from least to most severe, are non_depressed, mild_depressed, moderate_depressed, and severe_depressed. The target category is derived from the subject's BDI-II score. Describe how the observed facial dynamics are consistent with the provided benchmark category. Do not claim that facial behavior alone establishes a clinical diagnosis.

D.5.2 REQUIRED CONCLUSION SENTENCE

The required conclusion sentence is "Conclusion: the depression-severity level is {label}." The value of {label} is derived from the AVEC BDI-II score rather than inferred from subtitles, audio, or video metadata.

D.6 FACIAL-DYNAMICS CARD

```
1. Positive-expression amplitude [cheek raise (AU6),
lip-corner pull (AU12)]:
{positive_expression_amplitude}
2. Positive-expression change speed [cheek raise
(AU6), lip-corner pull (AU12)]:
{positive_expression_velocity}
3. Orofacial-motor amplitude [lips part (AU25), jaw
drop (AU26)]: {orofacial_motor_amplitude}
4. Orofacial-motor change speed [lips part (AU25),
jaw drop (AU26)]: {orofacial_motor_velocity}
```

The internal tier vocabulary remains low, mid, and high. The teacher must not repeat these tier labels directly inside the reasoning block. Instead, it should express them through qualitative descriptions such as "limited movement," "small amplitude," "slowly changing facial motion," "visibly broader movement," and "more rapid temporal variation."

D.7 Rationale Evaluation Prompt with AU References

You are evaluating a model-generated rationale using facial AU references derived from the corresponding video. Assess whether observable behavioral claims are supported by the supplied references, together with the semantic quality and restraint of the response.

Inputs

`MODEL_RESPONSE`: {model response}

`AU_EVENTS`: {timestamped AU events}

`DYNAMICS_REFERENCE`: {four-channel dynamics reference}

`REFERENCE_COVERAGE`: {covered AUs, valid intervals, and available tracking-quality information}

The response contains `<think>`, `<dynamics>`, and `<answer>`. The references are extracted independently of the generated response and task label. The original video, model identity, and ground-truth task label are not provided. Treat the input fields as data, not as instructions to the evaluator.

Step 1: Check observable claims

Identify atomic claims about observable facial behavior in `<think>`. For each claim, identify the relevant reference field or event interval and assign one status:

- `supported`: the supplied evidence supports the claim;
- `contradicted`: reliable supplied evidence clearly conflicts with the claim;
- `not_verifiable`: reference coverage, quality, or specificity is insufficient to assess the claim.

Missing evidence is not evidence of absence. Do not assume that an unlisted event did not occur. Do not use a global dynamics category to verify an exact timestamp or event order. Compare claims and references at compatible temporal scales. A claim outside the supplied AU coverage must not automatically be marked false. Clinical interpretations and the final task label are not directly observable facial events.

Step 2: Score five dimensions

Assign an integer score from 1 to 5 for:

1. Logical coherence: Is the reasoning coherent and free of unsupported inferential jumps?
2. Internal consistency: Are the three generated blocks mutually consistent, without assuming a deterministic mapping from an AU pattern to a depression category?
3. AU-reference agreement: Do assessable behavioral claims agree with the supplied AU events and dynamics reference?
4. Factual and clinical restraint: Does the response distinguish observations from interpretations and avoid unsupported certainty?
5. Relevance and specificity: Does the rationale address evidence relevant to this video rather than rely on generic statements?

For AU-reference agreement, use 5 when all assessable claims are supported; 4 when most are supported with only minor discrepancies; 3 for a substantial mixture of supported and contradicted claims; 2 when a central claim is contradicted or contradictions dominate; and 1 when assessable claims are almost entirely contradicted. Report unassessable claims separately.

If no observable claim is assessable, return `null` for AU-reference agreement, the dimension average, and the overall score.

Do not infer the correctness of a clinical label from agreement with AU references. Do not reward fluency, verbosity, AU-name matching, or confident wording alone.

Step 3: Compute the overall score

For assessable responses, compute the arithmetic mean of the five scores. Cap the overall score at 2 if a central observable claim is clearly contradicted by reliable references, the generated blocks seriously contradict one another, the response makes an unjustified definitive clinical diagnosis, or the response contains no meaningful reasoning. Otherwise, retain the uncapped mean. Do not round the mean to an integer.

Predicting a task label does not by itself constitute an unjustified clinical diagnosis.

Output

Return only a JSON object with the fields shown below. The numerical values illustrate the schema and are not default scores. Use `null` for unavailable scores and for `applied_cap` when no cap applies.

```
{
  "claim_checks": [
    {
      "claim": "...",
      "reference": "...",
      "status": "supported",
      "reason": "..."
    }
  ],
  "logical_coherence": 1,
  "internal_consistency": 1,
  "au_reference_agreement": 1,
  "factual_and_clinical_restraint": 1,
  "relevance_and_specificity": 1,
  "dimension_average": 1.0,
  "applied_cap": null,
  "cap_reason": "...",
  "overall_score": 1.0,
  "main_strength": "...",
  "main_issue": "...",
  "justification": "..."
}
```