VidEmo: Affective-Tree Reasoning for Emotion-Centric Video Foundation Models

Zhicheng Zhang [1] Weicheng Wang [1] Yongjie Zhu [3] Wenyu Qin [3] Pengfei Wan [3] Di Zhang [3] Jufeng Yang [1,2]

[1]Nankai University [2]Pengcheng Lab [3]Kuaishou Technology

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

- **High-Level emotion intelligence toward AGI requires explainable reasoning**: Human emotion in video is dynamic, open-set, and heavily context-dependent. Achieving AGI-level emotion understanding therefore requires models that not only perceive low-level cues but also construct ***transparent, step-wise*** affective rationales. However, even advanced VideoLLMs such as Gemini 2.0 reach only 26.3% accuracy on fine-grained sentiment analysis, underscoring the substantial gap in high-level emotional intelligence.
- **VidEmo: A Tree-Structured, Affective-Cue-Guided Reasoning Framework**: To close this gap, we propose VidEmo, a hierarchical reasoning model that decomposes emotion understanding into three stages: (1) ***attribute perception***, (2) ***expression analysis***, and (3) ***emotion reasoning***. Inspired by recent progress in reasoning-style models (e.g., R1),VidEmo performs structured, cue-guided reasoning, integrating appearance, micro-expressions, temporal dynamics, and scene context into coherent emotional interpretations.

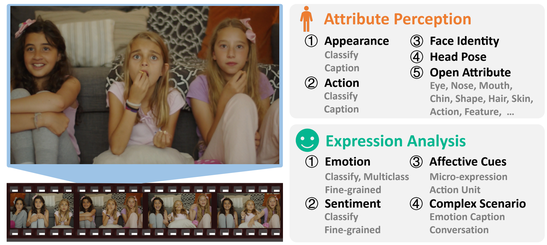

Figure 1. Selected examples of inputs and outputs obtained from . Apart from providing toolkits for basic attribute perception and expression analysis (top), extends the cognitive capacity and is able to generate fine-grained emotional captions with explainable rationale (bottom).

VidEmo: Video Emotion Foundation Models

Pipeline: To develop a family of emotion-centric video foundation models, we propose a comprehensive set of toolkits designed for the pre-training, post-training, and reasoning stages.

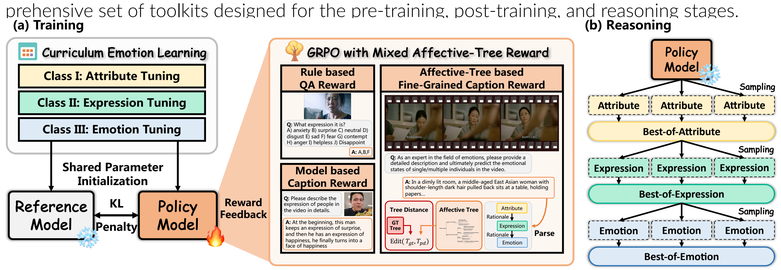

Figure 2. **Pipeline of VidEmo.** (a) Training: The model is trained using curriculum emotion learning, divided into three stages: attribute, expression, and emotion tuning. A reference model provides initial parameters, and a policy model is trained with reward feedback. (b) Reasoning: The policy model performs hierarchical reasoning by sampling from the best attributes, expressions, and emotions to generate the final emotional output.

Pre-training: Curriculum Emotion Learning

- **Curriculum Emotion Learning**: A progressive 3-stage curriculum that injects knowledge into base model: (1) attribute perception -> (2) expression analysis -> (3) emotion understanding. This reduces perplexity and stabilizes learning across difficulty levels.

Post-training: GRPO via Mixed Affective-Tree Reward

- **Starting from GRPO**: Formally, let q be a query, GRPO samples a group of outputs $\{o_i\}_{i=1}^{G}$ with the number of G from the old policy model $\pi_{\theta_{\text{old}}}$, and train a policy model by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\left(r_t(\theta)\hat{A}_{i,t},\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t} \right) \right] - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}),$$

- **Rule-based QA reward**: The model is evaluated on its ability to respond to emotion-related queries using predefined rules of Acc and F1 score.
- **Model-based Caption reward**: For the short caption of action, appearance, and emotion, we use a generative reward model to score the quality of captions generated by the model.
- **Affective Tree-based Fine-Grained Caption Reward**: Given a generated caption $\hat{o}$, we first parse it into a set of aspect–item pairs at three semantic levels: attribute ($\mathcal{A}$), expression ($\mathcal{E}$), and emotion ($\mathcal{M}$). These elements are organized into a three-level affective tree T_{pred}, where each node represents an extracted item and directed edges encode rationale:

$$\mathcal{A} \xrightarrow{\text{rationale}} \mathcal{E} \xrightarrow{\text{rationale}} \mathcal{M}. \tag{1}$$

We compare the predicted tree T_{pred} with gt tree T_{gt} using the tree edit distance $\text{Edit}(T_{\text{gt}}, T_{\text{pred}})$, which quantifies the minimal number of edit operations (insertions, deletions, substitutions) required to transform one tree into the other as reward R:

$$R = \exp\left(-\lambda \cdot \text{Edit}(T_{\text{gt}}, T_{\text{pred}})\right), \tag{2}$$

New Milestone: Our best model, *VidEmo*-T1, shows superior performance across 15 face perception tasks, surpassing advanced milestone (*Gemini 2.0: 5th Feb, 2025*) on 14 of 15 tasks

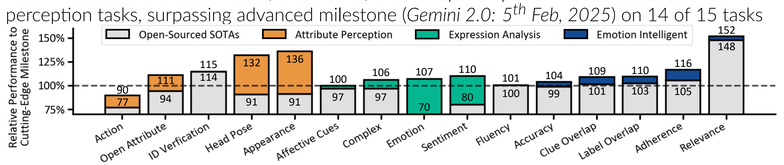

Figure 3. Results Overview.

SOTA comparison: We compare VidEmo with 18 leading VideoLLMs on 14 face attribute perception, 11 expression analysis and 6 fine-grained emotion understanding tasks. VidEmo achieves over a 16.3% and 14.2% improvement compared to existing open-source VideoLLMs on 1-3B and 7-8B scales.

Model	Size	Appearance		Action		ID	Head	AVG	Open Attribute										AVG
		Cls	Cap	Cls	Cap	Veri	Pose		Mout.	Nose	Hair	Chin	Shap.	Feat.	Acce.	Age	Gend.	Skin	Act.	
Closed MLLM API																				
Gemini 2.0 [82]	-	42.2	56.4	41.4	62.2	86.5	25.8	52.4	72.1	72.9	87.3	87.3	64.1	61.9	72.1	80.0	78.7	94.3	81.5	61.6	76.2
Claude 3.5 Sonnet [3]	-	39.1	58.8	35.7	61.1	63.6	22.0	46.7	70.8	67.6	84.4	77.8	68.9	56.1	54.7	77.3	79.5	93.3	48.9	61.4	67.6
Qwen-VL-MAX [63]	-	41.1	54.3	32.5	60.0	89.7	34.5	52.0	71.9	69.6	79.4	84.4	64.0	71.5	64.4	71.7	78.9	93.3	74.0	60.9	73.7
GPT-4o [51]	-	29.3	51.5	05.1	40.7	79.0	27.6	38.8	51.5	45.7	62.9	77.1	48.5	46.4	54.5	78.1	68.5	86.3	75.5	52.8	62.3
GPT-4o mini [51]	-	20.5	55.8	04.1	54.0	69.3	27.8	38.6	43.2	44.5	52.0	52.4	31.6	43.5	40.1	45.0	41.0	69.6	45.9	30.7	44.9
Open-sourced 1-3B Video MLLM																				
LLaVA-OV [25]	1B	36.3	34.9	00.2	47.6	50.3	14.4	25.6	41.0	49.7	36.0	50.9	46.7	32.0	39.0	48.1	29.8	87.6	20.1	61.9	45.2
InternVL2.5 [8]	1B	17.7	46.2	13.4	47.1	04.7	17.3	24.4	53.2	50.6	57.5	70.3	43.5	38.6	42.8	54.5	52.1	80.7	59.8	51.6	54.6
VideoLLaMA3 [86]	2B	36.8	05.2	48.5	89.2	26.6	33.4	55.9	51.5	52.0	73.7	47.1	52.7	55.0	85.4	59.3	52.1	95.3	60.1	22.4	59.4
mPLUG-Owl3 [81]	2B	16.0	45.4	13.8	**52.3**	76.4	07.7	38.3	54.4	61.4	55.4	71.4	43.7	45.4	52.3	60.8	39.8	91.6	58.7	48.5	56.9
Qwen2.5-VL [65]	3B	43.6	41.1	30.2	49.9	95.7	15.5	46.0	64.2	54.3	51.3	72.8	29.1	40.8	52.1	58.9	70.6	93.5	76.2	**62.6**	60.5
VidEmo-Base	3B	**57.0**	**47.9**	**37.7**	**47.9**	**100**	**90.7**	**84.9**	**84.7**	**82.7**	**94.0**	**85.2**	**75.9**	**80.8**	**75.0**	**83.4**	**84.0**	**95.0**	**88.8**	61.1	**82.8**
Open-sourced 7B+ Video MLLM																				
ShareGPT4Video [7]	8B	10.3	38.7	13.7	51.6	03.0	17.1	22.4	53.9	54.7	37.7	74.8	13.1	23.8	26.6	46.7	45.1	45.2	51.6	57.9	53.1	46.9
InternVL2.5 [8]	8B	36.9	38.7	17.4	49.8	61.2	15.5	36.6	56.3	59.0	55.0	72.2	52.4	36.4	52.7	61.5	60.0	76.9	57.6	59.3	58.3
LLaVA-N-Video [41]	7B	16.9	34.6	20.5	49.2	51.7	05.8	29.8	42.6	43.0	40.4	67.5	18.3	49.9	44.2	52.3	16.7	84.4	58.2	48.6	47.2
LLaVA-OV [25]	7B	05.6	37.3	12.2	46.4	97.2	19.8	36.4	53.0	47.6	50.4	64.0	35.3	32.2	49.9	55.8	72.9	94.6	47.5	59.0	55.2
VideoLLaMA3 [86]	7B	28.3	33.5	15.8	48.7	89.2	16.4	38.6	54.4	56.7	55.5	71.9	40.5	36.6	50.6	61.3	60.4	84.7	65.6	64.8	58.6
LLaVA-Video [91]	7B	14.5	38.2	14.1	46.0	88.7	20.3	37.0	55.0	57.0	50.2	72.6	21.3	29.4	39.3	68.3	79.4	93.0	62.5	63.8	61.5
mPLUG-Owl3 [81]	7B	34.3	41.6	21.3	55.1	66.4	21.1	40.0	55.0	56.2	48.0	70.5	39.9	48.1	50.5	58.8	61.8	89.7	62.0	60.8	58.4
Qwen2.5-VL [65]	7B	44.7	45.2	21.0	52.3	99.7	22.6	47.6	68.6	70.5	83.2	74.7	66.4	55.6	60.8	73.8	77.2	94.0	76.2	64.0	72.1
VidEmo-Base	7B	60.3	72.9	38.4	55.1	**99.8**	93.4	69.2	86.4	85.5	95.1	85.1	77.3	81.6	78.7	85.0	85.6	95.0	89.5	71.7	84.7
VidEmo-T1	7B	**64.8**	**73.1**	**41.4**	**57.4**	99.7	**96.7**	**72.1**	**88.2**	**87.8**	**95.6**	**87.8**	**79.2**	**82.0**	**80.8**	**85.7**	**86.9**	**97.0**	**90.4**	**74.3**	**86.2**

Model	Size	Emotion			AVG	Sentiment		AVG	Cues		AVG	Complex		AVG	Emotion Understanding					AVG
		Sin	Mul	Fine		Sin	Fine		Mic	AU		Cap	Conv		Lab	Cls	IA	RA	VTR	Flu	
Closed MLLM API																				
Gemini 2.0 [82]	-	45.3	42.2	23.6	37.0	45.0	26.0	35.5	16.2	36.1	26.1	42.0	50.6	46.3	51.0	52.6	58.2	60.2	66.8	92.1	63.5
Claude 3 Sonnet [3]	-	26.6	17.6	18.6	20.3	37.6	23.3	30.4	29.2	37.1	33.2	50.3	36.3	43.3	47.5	49.5	57.6	54.9	62.0	92.5	60.7
Qwen-VL-MAX [63]	-	29.3	25.8	24.6	26.6	32.6	21.3	27.0	11.7	24.0	17.9	46.4	49.3	47.8	47.9	50.7	57.4	56.6	63.6	92.0	61.4
GPT-4o [51]	-	20.6	19.0	04.3	14.6	29.6	18.0	23.8	29.6	02.3	16.0	38.1	43.6	40.8	30.0	35.4	40.8	43.2	50.9	87.4	48.0
GPT-4o mini [51]	-	19.6	20.2	07.3	15.7	29.9	16.6	23.3	21.9	08.9	15.4	40.7	47.6	44.2	35.5	39.0	46.2	46.7	54.4	91.0	52.1
Open-sourced 1-3B Video MLLM																				
LLaVA-OV [25]	1B	28.3	17.5	05.3	17.0	36.3	15.0	25.6	13.4	15.8	14.6	40.2	29.6	34.9	17.2	18.2	28.6	23.5	28.8	89.0	34.2
InternVL2.5 [8]	2B	23.0	16.1	08.3	15.8	31.6	15.6	23.6	09.7	07.4	08.5	43.6	26.0	34.8	40.1	42.9	51.6	49.6	56.4	90.8	55.2
VideoLLaMA3 [86]	2B	29.3	18.3	11.6	19.7	31.3	25.7	28.5	24.0	22.4	23.1	50.5	29.6	40.1	39.4	40.3	46.8	46.4	53.4	91.3	54.3
mPLUG-Owl3 [81]	2B	32.3	18.4	05.0	18.6	35.6	07.6	21.6	**28.0**	22.7	25.3	45.0	34.3	39.7	27.8	27.9	35.8	34.6	41.3	89.9	42.9
Qwen2.5-VL [65]	3B	36.0	23.7	09.3	23.0	36.6	20.3	28.5	12.1	23.0	17.6	44.2	40.3	42.2	38.2	40.9	49.7	49.2	56.3	91.4	54.3
VidEmo-Base	3B	**46.4**	**38.0**	**26.0**	**36.6**	**40.3**	**32.6**	**36.5**	**22.3**	**32.1**	**27.2**	**48.1**	**42.0**	**45.0**	**57.2**	**59.6**	**70.7**	**62.7**	**68.1**	**93.1**	**68.6**
Open-sourced 7B+ Video MLLM																				
ShareGPT4Video [7]	8B	07.6	06.0	04.6	06.1	38.0	14.3	26.1	09.7	01.4	05.6	46.2	32.3	39.2	16.1	18.8	34.4	21.8	26.0	91.1	34.7
InternVL2.5 [8]	8B	28.0	26.2	09.0	21.0	29.3	18.3	23.8	16.2	12.8	14.5	40.8	35.0	37.9	52.3	53.4	61.5	59.8	66.1	92.4	64.2
LLaVA-N-Video [41]	7B	24.3	23.7	10.6	19.5	39.0	14.0	26.5	10.9	13.3	12.1	44.1	39.0	41.5	33.7	33.2	43.3	38.8	45.2	90.9	47.5
LLaVA-OV [25]	7B	31.6	22.7	10.3	21.5	36.0	20.0	28.0	11.7	15.5	13.6	42.2	43.0	42.6	36.5	39.3	49.5	46.2	53.1	91.0	52.6
VideoLLaMA3 [86]	7B	27.6	23.9	10.6	20.7	31.0	19.3	25.1	08.9	10.8	09.8	33.2	36.0	34.6	42.2	44.4	53.1	52.9	59.5	89.5	56.9
LLaVA-Video [91]	7B	32.6	22.9	09.3	21.6	35.3	20.6	28.0	12.1	02.5	05.3	45.8	42.0	43.9	38.8	40.9	50.1	47.8	54.6	91.2	53.9
mPLUG-Owl3 [81]	7B	30.0	22.2	10.8	21.0	29.3	15.6	22.5	**21.5**	23.3	47.2	33.6	40.4	36.5	35.6	44.2	43.8	50.3	90.8	50.2	
Qwen2.5-VL [65]	7B	38.6	27.0	12.3	26.0	30.0	22.3	26.1	10.5	14.4	12.5	46.2	44.3	45.2	50.7	52.1	60.0	59.7	66.3	92.7	63.6
VidEmo-Base	7B	47.3	37.6	34.6	39.8	**42.3**	36.0	39.1	18.2	**34.2**	26.2	50.0	48.6	49.3	55.9	57.4	67.9	62.6	**68.3**	**92.8**	**67.5**
VidEmo-T1	7B	**49.7**	**38.8**	**38.6**	**41.3**	**42.3**	**37.5**	**39.9**	**20.4**	34.1	**27.3**	**50.7**	**52.9**	**51.8**	**59.3**	**61.2**	**68.1**	**65.9**	69.1	92.6	**69.3**

Emo-CFG: Emotion-Centric Fine-Grained Video Dataset

Data Curation: We curate Emo-CFG from 17 diverse video datasets, preserving rich meta information. We further obtain rationale annotations linking attribute, expression, and emotion. All data is rigorously filtered through a committee, ensuring high-quality emotion annotations.

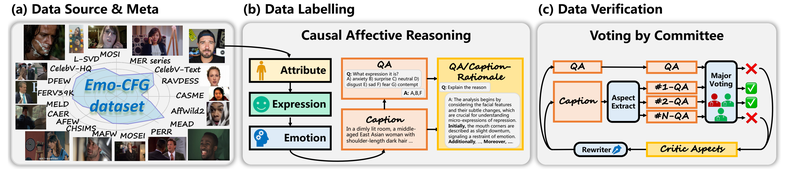

Figure 4. **Data Curation Pipeline of the dataset.** (a) The source of data from 17 datasets. (b) The illustration of data labeling steps. (c) The illustration of data verification loop.

Data Statistics: Emo-CFG covers three major task types with diverse facial views, video lengths, and affective annotations. Its large scale, rich label variety, and inclusion of fine-grained emotions and rationales make it substantially more comprehensive than existing emotion or video datasets.

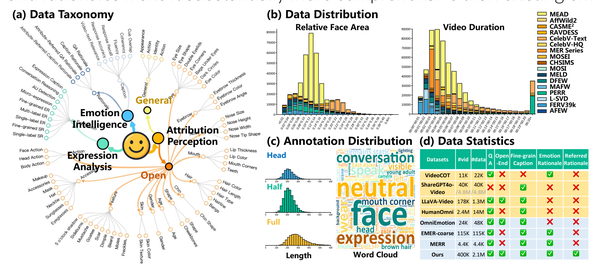

Figure 5. **Data Statistics of Emo-CFG.** (a) The data taxonomy from three types of face perception tasks. (b) The temporal and spatial distribution of video data. (c) The data label distribution and examples. (d) The comparison with other emotion and video datasets.

Datasets	#vid	#duration	Open-End	Fine-grained Caption	Emotion Rationale	Refered Rationale
VideoCOT	11K	22K	✓	✗	✗	✗
ShareGPT4o-Video	40K	40K	✗	✓	✗	✗
LLaVA-Video	178K	1.3M	✓	✓	✗	✗
HumanOmni	2.4M	16M	✓	✓	✗	✗
OmniEmotion	24K	48K	✓	✓	✗	✗
EMER-coarse	115K	115K	✓	✓	✓	✗
MERR	4.4K	4.4K	✓	✓	✓	✗
Ours	400K	2.1M	✓	✓	✓	✓

User Study: Emo-CFG is built to preserve high data quality even beyond 50K samples, a scale at which consistency becomes difficult for human-labeled datasets such as CelebV-Text. As a fully automated and systematically verified pipeline, Emo-CFG delivers reliable, large-scale emotional annotations across diverse video sources.

Dimension	Pairwise			#Vid	#Usr	Prefer	p-value
	Win	Tie	Loss				
Precision	964	204	82	50	25	95.5%	0.00021
Rationality	1082	87	81	50	25	92.1%	0.00015
Complementary	1172	23	55	50	25	93.0%	0.00008

Table 2. **User study between VidEmo and CelebV-Text.** We test the label quality on precision, rationality, and complementary through pairwise comparison with 50 videos and 25 users. All three dimensions show significant preference for VidEmo.

Conclusion

- We introduce VidEmo, a unified video emotion foundation model that integrates attribute perception, expression analysis, and high-level affect reasoning.
- We build Emo-CFG, a 2.1M-sample emotion-centric dataset with rich fine-grained annotations, establishing a comprehensive data infrastructure for community.

Contact
MODA
gloryzzc6@sina.com

Website

Project

Code