AffectAnchor: Grounding Affect with Multimodal Evidence via Retrieval Feedback

Anonymous ACL submission

Abstract

Multimodal emotion recognition (MER) requires reasoning over what is said, how it is spoken, and visible behavior, but many existing systems optimize only final emotion labels or generate rationales without direct evidence-level supervision. We introduce AffectAnchor, a retrieval-based framework that formulates affect prediction as evidence-to-description matching. Given a multimodal input, AffectAnchor generates linguistic, acoustic, and visual evidence and uses this evidence to rank structured affect descriptions. During training, a Qwen2.5-Omni-7B evidence generator is optimized with Group Relative Policy Optimization (GRPO) using retrieval-aware rewards over a fixed 73-concept meta affect pool, combining concept-coverage F1 and normalized discounted cumulative gain (NDCG). At inference time, the evidence generator and reranker remain fixed, and zero-shot target-task transfer is achieved by replacing the candidate affect description pool with the target taxonomy. On MER-UniBench, AffectAnchor reaches an average primary score of 82.08, outperforming AffectGPT-RL by 2.10 points and its Qwen2.5-Omni-7B backbone by 10.01 points. On OV-MERD evidence evaluation, it improves weighted evidence recall from 61.94% to 74.55% and modality coverage from 68.09% to 80.09%, while reducing unsupported-claim rate from 30.40% to 27.86%.

1 Introduction

Emotion perception is central to many applications such as human-computer interaction (Sinha et al., 2010) and educational assistance (Woolf et al., 2009). Although large language models have shown strong reasoning abilities for affect analysis (Wang et al., 2023; Kojima et al., 2022), emotion in real-world interaction is rarely expressed through language alone. It is shaped jointly by what a person says, how it is said, facial and bodily behavior, and the surrounding context. This has motivated multimodal emotion recognition (MER), which integrates linguistic, acoustic, and visual signals to infer human affective states.

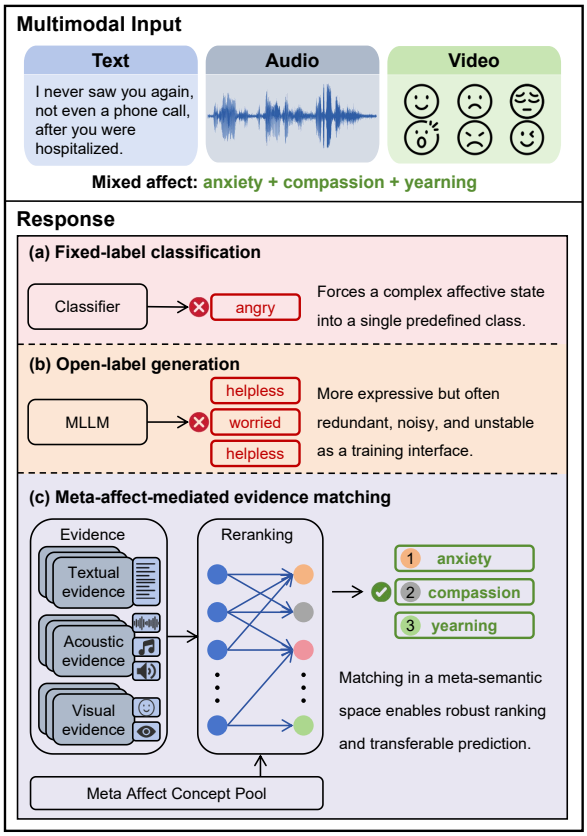

Figure 1: Comparison of emotion modeling paradigms. Our meta-affect-mediated evidence matching aligns multimodal affective evidence with candidate concepts, reducing label compression and open-label instability.

Despite this progress, most MER systems still treat affect understanding primarily as label prediction or explanation generation. Closed-set classification maps a complex and context-sensitive affective state into a small predefined taxonomy, which may discard fine-grained semantic distinctions (Russell, 1980; Lian et al., 2025c). Open-vocabulary approaches relax this restriction by al-

lowing richer affect labels or descriptions (Lian et al., 2025c), but they still mainly optimize the final affect output. Recent affect reasoning models, such as Emotion-LLaMA (Cheng et al., 2024), further generate natural-language rationales, yet these rationales are not necessarily trained or evaluated according to whether they provide evidence that is semantically aligned with the target affect. As a result, a model may produce a plausible label or a fluent explanation without identifying the multimodal cues that make the affective interpretation recoverable (Huang et al., 2023).

We argue that multimodal affect understanding should be grounded in evidence before being projected to a task-specific label space. We introduce multimodal evidence-based affect grounding, as illustrated in Figure 1. Instead of directly asking an MLLM to output an emotion label, we ask whether it can generate affective evidence that is sufficiently specific to support the identification of a target person's affective state (Luo and Shakhnarovich, 2017). We formulate this grounding process as retrieval: generated affective evidence serves as the query, candidate affect descriptions serve as documents, and affect grounding is performed through semantic matching (Nogueira and Cho, 2019). This formulation separates evidence reasoning from final label selection, allowing affective semantics to be represented, compared, and ranked before being mapped to a downstream taxonomy.

Building on the above task definition, we propose AffectAnchor, a framework for improving the affective evidence reasoning ability of MLLMs. AffectAnchor constructs a structured meta affect concept pool in which each affect concept is represented by an enriched textual description rather than a bare label. Given multimodal input, the model generates linguistic, acoustic, and visual evidence, and the evidence is used to retrieve semantically compatible affect concepts. We optimize this evidence generation process with reinforcement learning (Schulman et al., 2017; Ouyang et al., 2022), using retrieval-aware rewards that combine F1 (Powers, 2020) and NDCG (Järvelin and Kekäläinen, 2002; Valizadegan et al., 2009). The reward encourages generated evidence to cover relevant affect concepts while preserving their relative ranking among semantically close candidates. In this way, supervision is tied directly to grounding performance rather than only to final labels or free-form explanations.

A further advantage of this retrieval view is that it naturally supports zero-shot transfer across affect taxonomies. Because AffectAnchor learns to produce evidence that matches textual affect descriptions, the learned grounding process is not bound to a single label inventory. At inference time, the same evidence generator and reranking mechanism can be reused with a different target description set, such as polarity labels, basic emotion categories, fine-grained affect concepts, or intensity descriptions (Radford et al., 2021; Lian et al., 2025c). Thus, downstream adaptation is performed by replacing the target-side affect description pool rather than by retraining the model for each taxonomy.

In summary, our core contributions are:

- We propose multimodal evidence-based affect grounding as a query-document matching problem, offering a unified retrieval-based perspective on affect grounding.
- We propose AffectAnchor, which uses reinforcement learning with retrieval rewards based on F1 and NDCG to explicitly improve the affective evidence reasoning and evidence generation abilities of MLLMs.
- We demonstrate that higher-quality affective evidence supports zero-shot transfer by simply replacing affect descriptions, yielding stronger generalization across different affect targets.

2 Related Work

Multimodal Emotion Recognition (MER). MER has traditionally been formulated as closed-set classification on benchmarks such as IEMOCAP (Busso et al., 2008), CMU-MOSEI (Zadeh et al., 2018), and OMG-Emotion (Barros et al., 2018), often using multimodal fusion architectures (Zadeh et al., 2017; Tsai et al., 2019; Hazarika et al., 2020). While effective within fixed taxonomies, these methods compress nuanced and context-dependent affect into predefined labels and generalize poorly across tasks, despite psychological accounts that view affect as dimensional or structured around basic emotions (Russell, 1980; Ekman, 1992). Recent open-vocabulary methods such as OV-MER (Lian et al., 2025c) relax label constraints, but affect remains semantically continuous and generated rationales may not faithfully reflect the evidence used by the model

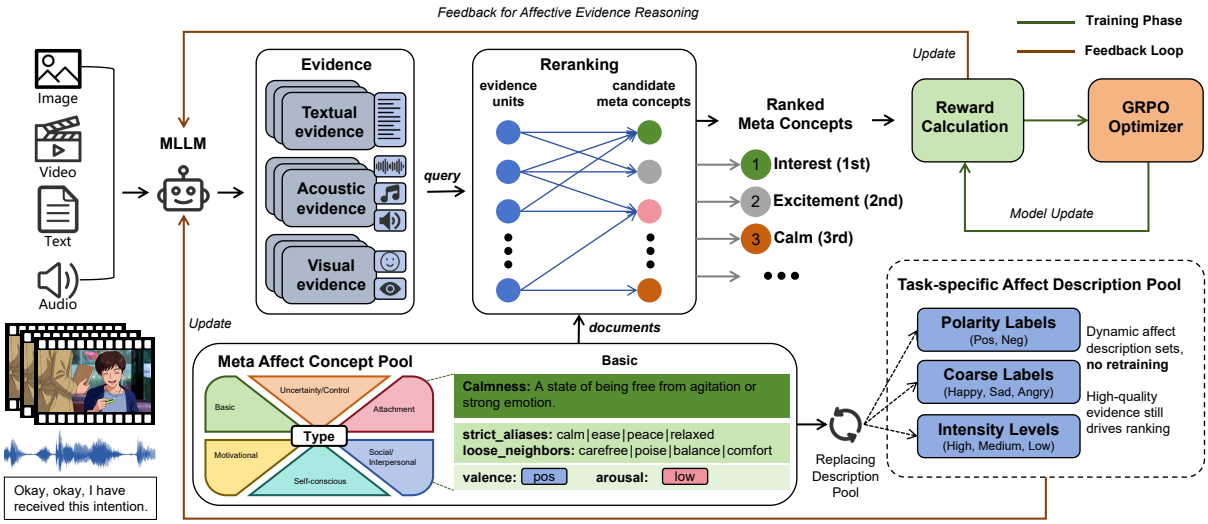

Figure 2: Overview of **AffectAnchor**. Multimodal inputs are converted into affective evidence, which is used as a query for evidence-conditioned reranking. Training optimizes this evidence against a structured meta affect concept pool with retrieval-based GRPO rewards. For transfer, the candidate document pool is replaced with task-specific affect descriptions while the evidence generator and reranker remain fixed.

(Cowen and Keltner, 2017; Huang et al., 2023). In contrast, we introduce a canonical intermediate affective semantic space and ground affect through evidence-to-concept matching, enabling flexible projection to downstream label sets.

Retrieval-based Semantic Matching. Retrieval systems provide a natural interface for matching queries with textual candidate descriptions, especially when the candidate space is open-ended or meaning-defined rather than tied to fixed label IDs (Karpukhin et al., 2020; Thakur et al., 2021; Nogueira and Cho, 2019). This separation between evidence representation and target selection is well suited to affect understanding, where labels are semantically close, context-dependent, and uneven in granularity. Following work on description-based and zero-shot semantic matching (Yin et al., 2019; Basile et al., 2022), AffectAnchor treats generated multimodal evidence as the query and affect descriptions as documents. This grounds prediction in evidence-description compatibility and supports transfer by replacing the candidate description set.

3 Methods

As shown in Figure 2, AffectAnchor formulates multimodal affect grounding as evidence-to-description ranking. The model first generates affective evidence from linguistic, acoustic, and visual cues, and a reranker then scores this evidence against textual affect descriptions. During training, the retrieval space is a structured meta affect concept pool, which provides concept-level supervision for improving evidence generation. At inference time, the same evidence generator and reranker are reused with task-specific affect descriptions, allowing zero-shot transfer across different affect taxonomies.

3.1 Affect Grounding with a Meta Affect Concept Pool

Multimodal affect grounding requires a model to connect observed affective cues with the affect states they support. Useful evidence should therefore contain linguistic, acoustic, and visual cues that make the target affect state identifiable. We use a reranker as the scoring module for this matching process. Given generated affective evidence as a query and an affect description as a document, the reranker estimates their semantic compatibility and assigns a matching score to each candidate description (Nogueira and Cho, 2019). The resulting scores induce a ranked list of affect candidates, so the final grounding depends on whether the generated evidence provides enough information for the reranker to identify the supported affect state. To make this matching process more informative, we define the candidate affect space at the concept level rather than using bare label names alone. An affect label provides a compact name for a state, while a concept-level entry makes its meaning explicit through a canonical label, definition, aliases, neighboring concepts, and affec-

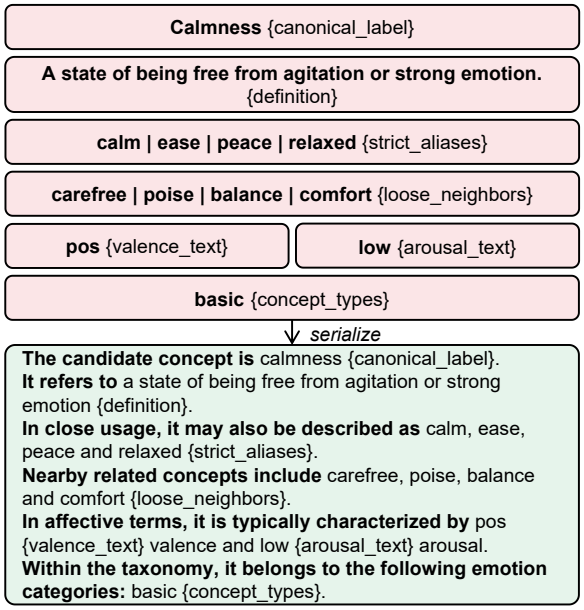

Figure 3: Serialization of a meta affect concept entry into a natural-language candidate document for evidence-to-concept reranking.

tive attributes. Such enriched descriptions give the reranker more semantic context for comparing generated evidence with candidate affect meanings. We therefore build a meta affect concept pool, which turns affect labels into structured retrieval targets and provides a stable semantic layer between multimodal evidence and downstream affect spaces.

We construct the meta affect concept pool from EMER (Lian et al., 2024) through a two-stage normalization process. In Stage I, we perform canonicalization and seed concept extraction: starting from 248 raw labels, we merge semantically equivalent entries, normalize canonical names and aliases, and remove non-affective or ill-defined items through manual verification, yielding 33 seed concepts. In Stage II, we conduct semantic expansion and consolidation: the seed concepts are expanded through LLM-based semantic augmentation and then filtered and merged through iterative human review. This process results in 73 canonical affect concepts. Full construction details and intermediate statistics are provided in Appendix A. Prompt summaries are provided in Appendix F.

The final meta affect concept pool contains 73 canonical concepts, each represented by a structured semantic entry consisting of a concept identifier, canonical label, definition, strict aliases, loose neighbors, valence, arousal, and affect type. As shown in Figure 3, each structured entry is serialized into a natural-language candidate document for reranking. These entries serve as the document space for reranking, support stable evidence-to-concept matching, and enable consistent projection to different downstream affect targets.

3.2 AffectAnchor: RL for Affective Evidence Reasoning

Given the concept-level grounding setup above, we train the MLLM to generate evidence that supports affect concept retrieval. AffectAnchor uses retrieval-aware rewards computed over the meta affect concept pool, so the learning signal is tied to how well the generated evidence identifies relevant affect concepts. This encourages the model to produce evidence that can be directly used for downstream affect grounding.

Reward Design. For an input x, the policy generates an output containing affective evidence e and an optional label prediction. Following the serialization scheme in Figure 3, only the generated evidence is used as the query to the reranker. The reranker scores each concept document in the meta affect concept pool and produces an evidence-conditioned ranking of candidate concepts.

We use two complementary retrieval rewards: a rank-insensitive concept coverage reward and a rank-aware NDCG reward. Let $G_x = (g_1, \ldots, g_m)$ denote the reference ordered list of relevant concepts for x, where $m = 5$ in the main setting, and let $\mathcal{G}_x$ be the corresponding set of canonical labels. Let $\hat{G}_K(e) = (\hat{g}_1, \ldots, \hat{g}_K)$ denote the top-K concepts retrieved using evidence e, and let $\mathcal{P}_K(e)$ be the corresponding set of canonical labels.

We compute a set-level F1 reward:

$$R_{\mathrm{F1}}(x, e) = \frac{2|\mathcal{P}_K(e) \cap \mathcal{G}_x|}{|\mathcal{P}_K(e)| + |\mathcal{G}_x|}. \tag{1}$$

In the main setting, $K = m = 5$, so this term is equivalent to normalized top-5 overlap. It ignores rank order and measures whether the generated evidence covers the reference affect concept neighborhood.

To reward ranking quality, we compute NDCG with linear gain. The reference list defines a graded relevance function:

$$\mathrm{rel}_x(c) = \begin{cases} m - i + 1, & \text{if } c = g_i \in G_x, \\ 0, & \text{if } c \notin \mathcal{G}_x. \end{cases} \tag{2}$$

Given the predicted ranking $\hat{G}_K(e)$, we compute:

$$\mathrm{DCG@}K(x,e) = \sum_{j=1}^{K} \frac{\mathrm{rel}_x(\hat{g}_j)}{\log_2(j+1)}, \quad (3)$$

and the ideal discounted cumulative gain is:

$$\mathrm{IDCG@}K(x) = \sum_{j=1}^{\min(K,m)} \frac{m-j+1}{\log_2(j+1)}. \quad (4)$$

The normalized ranking reward is:

$$R_{\mathrm{N@}K}(x,e) = \frac{\mathrm{DCG@}K(x,e)}{\mathrm{IDCG@}K(x)}. \quad (5)$$

This reward uses rank-derived grades rather than raw reranker scores, avoiding sample-level score calibration issues while preserving the relative order of reference concepts.

The final retrieval reward is:

$$R_{\mathrm{ret}}(x,e) = \lambda_{\mathrm{F1}} R_{\mathrm{F1}}(x,e) + \lambda_{\mathrm{N}} R_{\mathrm{N@5}}(x,e), \quad (6)$$

where λ_{F1} and λ_{N} balance concept coverage and ranking quality.

Group Relative Policy Optimization. We optimize the policy with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each input, we sample multiple candidate outputs, compute their retrieval rewards, and normalize the rewards within the group to obtain relative advantages. The policy is then updated to favor outputs with higher retrieval rewards under a clipped policy-gradient objective. Following the standard GRPO formulation, the objective can include KL regularization against a reference policy; in our main experiments, we set the KL coefficient to zero, as detailed in Appendix C. AffectAnchor uses GRPO to optimize affective evidence generation according to retrieval quality in the meta affect concept space. The optimized evidence can then be matched with different downstream affect description sets, supporting zero-shot transfer across target taxonomies.

3.3 Zero-shot Transfer via Affect Description Replacement

Because AffectAnchor optimizes evidence generation in a shared affect-semantic space, the learned grounding policy is not tied to a single task-specific label taxonomy. For a new task, we keep the evidence generator and reranker fixed, and replace only the target-side affect description set. Each target label is represented by a textual description, and the generated evidence is matched against these descriptions with the same reranker used for concept grounding. The final prediction is the label whose description receives the highest compatibility score.

This design transfers affect knowledge by reusing the evidence-generation ability learned from the meta affect concept space, while adapting to a new target taxonomy through textual label descriptions. No task-specific fine-tuning or target-label supervision is required. The same mechanism can be applied to different affect prediction settings, including polarity classification, basic emotion recognition, and fine-grained or open-vocabulary emotion recognition.

In our experiments, this protocol is instantiated on MER-UniBench. The model is not fine-tuned on the target benchmarks, and each benchmark is accessed only through its target label descriptions at inference time. We therefore treat MER-UniBench performance as zero-shot transfer across affect taxonomies rather than supervised adaptation to a specific dataset.

4 Experiments

We evaluate whether retrieval-aligned optimization improves multimodal affect grounding by improving affective evidence reasoning. We compare AffectAnchor with representative multimodal baselines on MER-UniBench, conduct zero-shot evidence evaluation on EMER, and provide qualitative analyses and ablations to examine the effects of the meta affect concept pool, retrieval-based reward, and RL-based optimization. All experiments are conducted on NVIDIA A100-80G GPUs.

4.1 Experimental Setup

Before evaluation, we construct the meta affect concept pool from the EMER label vocabulary and keep it fixed. Target taxonomies are used only as inference-time label descriptions; no target instances, annotations, validation labels, predictions, or feedback are used for training, reward construction, prompt tuning, hyperparameter/model selection, or pool revision. Thus, the setting is zero-shot with respect to target-task data and supervision, while using an external affect-vocabulary prior as the intermediate concept space. Following the MER-UniBench protocol (Lian et al., 2025a), we audit MER-caption+–target overlap across sample,

Model	Prec.	Wtd. Recall	Halluc. Rate ↓	Mod. Cov.
E-LLaMA	53.31	44.71	39.22	50.41
AffectGPT	52.74	49.95	43.78	51.96
Qwen-Omni	62.33	61.94	30.40	68.09
AffectAnchor	**67.55**	**74.55**	**27.86**	**80.09**

Table 1: Atomic evidence quality on OV-MERD under zero-shot evaluation. Scores are percentages (%).

source, subtitle, description, speaker/session, and label-vocabulary fields; details are in Appendix B.

Training Data. We train AffectAnchor on a reranked version of MER-caption+ (Lian et al., 2025b), which contains 31K multimodal samples with video, audio, subtitles, rich emotion descriptions, and open-vocabulary affect labels. Training optimizes evidence-induced concept ranking in the meta affect concept space rather than final dataset-specific labels.

MER-UniBench Evaluation. We evaluate on the reranking-mapped MER-UniBench (Lian et al., 2025b), covering sentiment, basic emotion, and fine-grained emotion recognition. At inference time, the generator and reranker are fixed, and only the target affect-description pool is replaced. We use the official metrics: WAF for sentiment datasets, HIT for basic emotion datasets, and grouped F_S for fine-grained and open-vocabulary tasks, with the dataset-wise mean as the overall score. Baselines include general-purpose multimodal models, PandaGPT and R1-Omni (Su et al., 2023; Zhao et al., 2025), and affect-oriented models, including Emotion-LLaMA, AffectGPT, Nano-EmoX, AffectGPT-RL, and EmotionLLaMAv2 (Cheng et al., 2024; Lian et al., 2025a; Huang et al., 2026; Lian et al., 2026; Peng et al., 2026).

4.2 Implementation Details

AffectAnchor is initialized from Qwen2.5-Omni-7B (Xu et al., 2025), denoted as Qwen-Omni in the tables. We freeze the visual encoder and train LoRA adapters (Hu et al., 2022) directly with GRPO, without SFT warm-up. A frozen Qwen3-Reranker-8B (Zhang et al., 2025) computes rewards over the fixed meta affect concept pool, using generated evidence as the query and top-5 reranked MER-caption+ concepts as references. We use $R_{\text{ret}} = R_{\text{F1}} + R_{\text{N@5}}$ with equal weights. Training runs for one epoch with four completions per prompt, learning rate 1×10^{-5}, and group-wise reward normalization. Training uses four NVIDIA A100-80G GPUs for approximately two days. Further details and baseline settings are provided in Appendix C.

4.3 Zero-shot Evidence Evaluation

To examine whether final prediction gains stem from better intermediate grounding, we conduct a zero-shot evidence evaluation on OV-MERD. GPT-5.4 decomposes gold descriptions into hidden atomic evidence units and model reasoning responses into generated claims. Each unit represents a minimal affect-relevant cue from text, audio, vision, or cross-modal context.

We compare generated units with hidden references using evidence precision, hallucination rate, weighted evidence recall, and modality coverage. Evidence precision measures the supported portion of generated claims:

$$\text{Prec.} = \frac{1}{N_c} \sum_{c=1}^{N_c} \max(v(c), 0). \tag{7}$$

Hallucination rate measures unsupported or contradicted claims:

$$\text{Halluc.Rate} = \frac{N_{\text{unsup}} + N_{\text{contra}}}{N_c}. \tag{8}$$

Weighted evidence recall measures covered reference evidence:

$$\text{Wtd.Recall} = \frac{\sum_{e \in E_{\text{covered}}} w(e)}{\sum_{e \in E_{\text{all}}} w(e)}. \tag{9}$$

Modality coverage measures whether generated evidence covers the required text, audio, visual, and cross-modal cues. Matching details and prompt templates are provided in Appendix D and Appendix F.

As shown in Table 1, AffectAnchor improves both evidential correctness and coverage while reducing unsupported cues. The gain in weighted recall suggests that retrieval feedback helps the model recover a broader affect-relevant evidence neighborhood, rather than focusing only on isolated salient cues. Appendix D reports human and open-judge validation, confirming that the system-level ranking is stable.

4.4 Zero-shot Transfer on MER-UniBench

Table 2 reports zero-shot target-task transfer on MER-UniBench. AffectAnchor obtains the highest average score in our evaluation, exceeding the

Model	MOSI	MOSEI	SIMS	S-v2	MER23	MER24	MELD	IEMO	OV+	Avg.
PandaGPT	61.92	67.61	68.38	67.23	40.21	51.89	37.88	44.04	37.12	52.92
R1-Omni	58.02	56.48	71.82	68.58	64.17	67.43	43.20	51.58	55.24	59.61
E-LLaMA	66.13	67.66	78.32	77.23	59.38	73.62	46.76	55.47	52.97	64.17
AffectGPT	81.30	80.90	**88.49**	86.18	78.54	78.80	55.65	60.54	62.52	74.77
Nano-EmoX	76.82	79.81	86.25	84.76	79.09	77.94	56.55	60.12	64.75	74.01
AffectGPT-RL	79.65	80.18	87.26	85.75	84.51	**93.13**	66.71	74.26	**68.39**	79.98
E-LLaMAv2	**86.28**	**87.69**	87.72	86.02	77.28	86.90	51.32	**84.05**	62.90	78.91
AffectAnchor	85.44	85.95	88.06	**87.78**	**89.70**	91.12	**68.68**	75.26	66.71	**82.08**

Table 2: Zero-shot cross-taxonomy transfer results on MER-UniBench. MOSI, MOSEI, SIMS, S-v2, MER23, MER24, IEMO, and OV+ denote CMU-MOSI, CMU-MOSEI, CH-SIMS, CH-SIMS v2, MER2023, MER2024, IEMOCAP, and OV-MERD+, respectively. E-LLaMA denotes Emotion-LLaMA. Each column reports the official primary metric for the corresponding dataset, and Avg. is the unweighted mean over datasets. Complete metric results are provided in Appendix E. **Bold** and underline indicate the best and second-best results.

Variant	Avg.	Δ
Qwen-Omni (no RL)	72.07	–
Ours w/o NDCG Reward	79.02	+6.95
Ours w/o F1 Reward	80.09	+8.02
Ours	**82.08**	**+10.01**

Table 3: Reward ablation under the fixed evidence-to-description pipeline.

Model	Reranker	Avg.
Qwen-Omni	Qwen3-Reranker-8B	72.07
AffectAnchor	Qwen3-Reranker-8B	**82.08**
Qwen-Omni	BGE-Reranker-v2-M3	71.35
AffectAnchor	BGE-Reranker-v2-M3	**80.12**

Table 5: Cross-reranker transfer of generated evidence.

Variant	Avg.	Δ
Ours w/ 73 Canonical Labels	80.47	+8.40
Ours w/ 238 Label Variants	80.34	+8.26
Ours w/ Structured Meta Pool	**82.08**	**+10.01**

Table 4: Effect of candidate description format.

published AffectGPT-RL score by 2.10 points. As MER-UniBench baselines are commonly reported as single-score results, we treat these comparisons as point estimates. The improvement is most pronounced on datasets whose label spaces differ substantially from the training-time meta concept space, such as SIMS v2, MER2023, and MELD. This suggests that optimizing evidence for retrieval can help the model preserve affect-relevant cues that remain useful after the candidate description pool is replaced.

AffectAnchor does not dominate every individual benchmark: Emotion-LLaMAv2 remains stronger on several sentiment and IEMOCAP settings, while AffectGPT-RL performs best on MER2024 and OV-MERD+. Rather than relying on gains from a single benchmark family, AffectAnchor shows a more balanced transfer profile across sentiment, basic emotion, and fine-grained affect taxonomies. The detailed transfer protocol is provided in Appendix B, where we specify that target benchmark instances, annotations, validation labels, predictions, and evaluation feedback are not used for training or model selection.

4.5 Ablation Studies

We examine whether the gains come from retrieval-reward optimization, the semantic form of candidate descriptions, or dependence on a particular reranker. Unless otherwise stated, target descriptions, decoding settings, and evaluation protocols are fixed. The Qwen-Omni baseline uses the same evidence-to-description pipeline without RL, separating the effect of optimization from prompting and reranking.

Table 3 shows that neither reward term alone accounts for the full improvement. Coverage-oriented F1 and rank-sensitive NDCG both yield substantial gains over the no-RL baseline, while their combination performs best. This suggests that the reward terms impose complementary pressures: F1 broadens the recovered affect neighborhood, whereas NDCG encourages finer ordering among semantically close concepts.

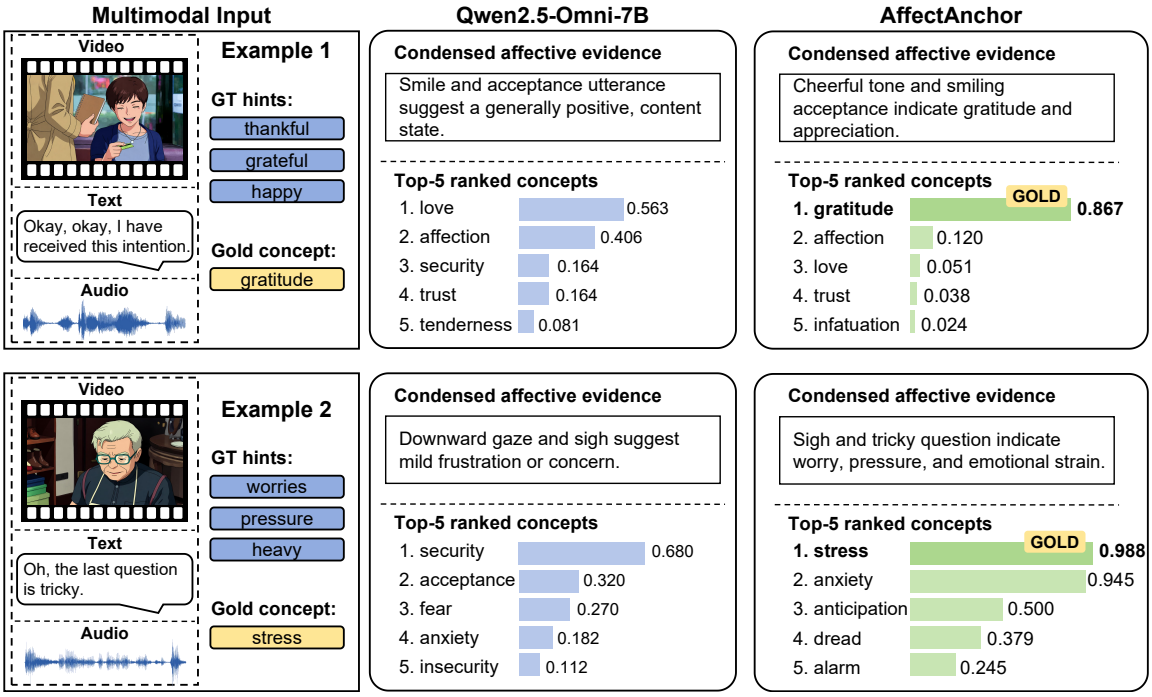

Figure 4: Qualitative visualization of evidence-based affect grounding. Visual frames are anonymized for privacy and copyright reasons, while utterances, ranking scores, and affect concepts are derived from real evaluation samples. Generated responses are shortened for readability while preserving the key multimodal cues and affective inferences.

The representation ablation in Table 4 further indicates that the candidate space is not merely a larger label inventory. Expanded surface variants do not improve over canonical labels, while structured descriptions bring the strongest transfer performance. The gain therefore appears to come from semantic grounding cues such as definitions, aliases, neighboring concepts, and affective attributes, rather than lexical expansion alone.

Table 5 examines whether the learned evidence is tied to the Qwen3 scorer used in training. Switching to BGE (Li et al., 2023; Chen et al., 2024) lowers absolute scores for both the no-RL baseline and AffectAnchor, reflecting differences in reranker calibration and semantic granularity. Nevertheless, AffectAnchor preserves a large margin under BGE, suggesting that retrieval-reward training improves the generated evidence in a form that transfers across reranker families, although final performance remains sensitive to reranker quality.

4.6 Qualitative Analysis

Complementing the quantitative results, Figure 4 illustrates grounding behavior at the example level. Retrieval-aligned training appears to shift evidence from broad affective impressions toward cues that are easier to verify in the input, including lexical content, vocal expression, and visible behavior. Such specificity is central to the retrieval formulation, where downstream decisions depend on evidence that discriminates among semantically neighboring affect concepts. In these cases, AffectAnchor provides clearer cues for the reranker, which helps promote the reference concept from outside the top-5 candidates to the highest rank. These examples provide qualitative support for the mechanism suggested by the aggregate results.

5 Conclusion

We presented AffectAnchor, a retrieval-based framework that grounds multimodal affect prediction in generated linguistic, acoustic, and visual evidence. By aligning evidence with structured affect descriptions through retrieval-aware GRPO rewards, AffectAnchor enables zero-shot cross-taxonomy transfer by replacing only the target description pool. On MER-UniBench, it achieves an average primary score of 82.08, outperforming AffectGPT-RL by 2.10 points, while OV-MERD evidence evaluation and ablations show improved cue coverage and reduced unsupported evidence. These results demonstrate that evidence-to-description grounding offers an effective semantic interface for transferable multimodal emotion recognition.

Limitations

AffectAnchor uses a compact affect-semantic interface rather than an exhaustive emotion ontology. This design supports stable retrieval and cross-taxonomy transfer, but the current 73-concept pool is primarily English-centered and may need additional aliases, descriptions, or concept variants for specialized domains, multilingual settings, and culturally specific affective expressions.

The zero-shot transfer step depends on the wording and granularity of target affect descriptions. We keep target descriptions fixed and do not tune them on target validation sets, but future work could study description calibration, automatic description selection, or richer affect taxonomies.

Our current implementation uses a modular generator-reranker pipeline. This modularity makes the evidence-to-description mechanism transparent and easy to analyze, but it introduces additional inference cost compared with a single classifier. Distillation, smaller rerankers, or cached candidate representations could improve efficiency.

Finally, our experiments focus on established benchmark-style multimodal clips. Extending the same evidence-grounded formulation to longer conversations, streaming interactions, or domain-specific deployments remains a direction for future work.

Ethical Considerations

This work studies multimodal affect understanding from language, speech, and visual behavior. Such data can be sensitive because it may contain personally identifying information and may be interpreted as evidence about a person's affective state. Our work is intended for research on evidence-grounded apparent affect under benchmark settings, not for measuring a person's true internal emotional state, mental health condition, intention, or personality.

We use existing research benchmarks and do not collect new primary multimodal recordings, personal videos, speech recordings, or user interaction data. All experiments follow the stated licenses, access conditions, and intended research uses of the corresponding datasets, and we do not redistribute original media. Dataset consent and license conditions are inherited from the source benchmarks and do not authorize new deployment contexts.

We recruited expert reviewers and annotators for concept-pool verification and evidence-evaluation validation. Participants were informed of the research purpose, expected workload, voluntary nature of participation, compensation policy, and data-use policy before the task, and they could withdraw at any time. We collected only minimal information needed for responsible reporting, such as self-reported expertise and country or region of residence, and did not collect sensitive personal attributes. For validation studies, system identities were anonymized before annotation.

For qualitative analysis, visual frames are anonymized, and examples are shortened to focus on affective cues relevant to model behavior. Any release of models, prompts, annotations, evaluation scripts, or the meta affect concept pool should be accompanied by documentation describing the intended research use, source artifacts, dataset licenses, and appropriate usage boundaries.

Automatic affect analysis systems may reflect biases present in training data, annotation practices, or cultural assumptions about emotional expression. Benchmark performance does not establish demographic fairness, cross-cultural validity, multilingual robustness, or reliability under domain shift. Systems based on this work should not be used for individual-level emotion inference about identifiable people, nor as standalone tools for high-stakes applications such as clinical diagnosis, hiring, surveillance, education assessment, law enforcement, insurance, or credit decisions. Practical deployment, if considered, should require informed consent, clear disclosure, data minimization, human oversight, and careful evaluation across demographic, cultural, linguistic, and domain-specific groups.

References

Pablo V. A. Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. 2018. The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pages 1–7. IEEE.

Angelo Basile, Marc Franco-Salvador, and Paolo Rosso. 2022. Unsupervised ranking and aggregation of label descriptions for zero-shot classifiers. In *Natural Language Processing and Information Systems - 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15-17, 2022, Proceedings*, Lecture Notes in Computer Science, pages 119–126. Springer.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 2318–2335. Association for Computational Linguistics.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G. Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jiahao Huang, Fengyan Lin, Xuechao Yang, Chen Feng, Kexin Zhu, Xu Yang, and Zhide Chen. 2026. Nano-emox: Unifying multimodal emotional intelligence from perception to empathy. *CoRR*, abs/2603.02123.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. Can large language models explain themselves? A study of llm-generated self-explanations. *CoRR*, abs/2310.11207.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models A better foundation for dense retrieval. *CoRR*, abs/2312.15503.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. 2025a. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, Ziyang Ma, Xiaojiang Peng, Xie Chen, Ya Li, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2025b. MER 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pages 13837–13842. ACM.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. 2025c. OV-MER: towards open-vocabulary multimodal emotion recognition. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, Lan Chen, Shan Liang, Ya Li, Jiangyan Yi, Bin Liu, and Jianhua Tao. 2024. Explainable multimodal emotion recognition. *Preprint*, arXiv:2306.15401.

Zheng Lian, Fan Zhang, Lan Chen, Yazhou Zhang, Rui Liu, Jinyang Wu, Haoyu Chen, Xiaobai Li, Xiaojiang Peng, Bin He, and Jianhua Tao. 2026. Affectgpt-rl: Revealing roles of reinforcement learning in open-vocabulary emotion recognition. *Preprint*, arXiv:2605.06126.

Ruotian Luo and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. In

2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 3125–3134. IEEE Computer Society.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xiaojiang Peng, Jingyi Chen, Zebang Cheng, Bao Peng, Fengyi Wu, Yifei Dong, Shuyuan Tu, Qiyu Hu, Huiting Huang, Yuxiang Lin, Jun-Yan He, Kai Wang, Zheng Lian, and Zhi-Qi Cheng. 2026. Emotion-llamav2 and mmeverse: A new framework and benchmark for multimodal emotion understanding. *Preprint*, arXiv:2601.16449.

David M. W. Powers. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *CoRR*, abs/2010.16061.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 8748–8763. PMLR.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Gaurav Sinha, Rahul Shahi, and Mani Shankar. 2010. Human computer interaction. In *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, pages 1–4. IEEE.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6558–6569. Association for Computational Linguistics.

Hamed Valizadegan, Rong Jin, Ruofei Zhang, and Jianchang Mao. 2009. Learning to rank by optimizing NDCG measure. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 1883–1891. Curran Associates, Inc.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. 2023. Emotional intelligence of large language models. *CoRR*, abs/2307.09042.

Beverly Park Woolf, Winslow Burleson, Ivon Arroyo, Toby Dragon, David G. Cooper, and Rosalind W. Picard. 2009. Affect-aware tutors: recognising and responding to student affect. *Int. J. Learn. Technol.*, 4(3/4):129–164.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3912–3921. Association for Computational Linguistics.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1103–1114. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset

and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2236–2246. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *CoRR*, abs/2503.05379.

A Meta Affect Concept Pool Construction

This appendix describes how we construct the meta affect concept pool used by AffectAnchor. The goal of the pool is not to preserve all raw open-vocabulary labels, but to build a stable intermediate semantic space for evidence-to-affect grounding. Starting from the 248 fine-grained labels in EMER, we construct the pool through a two-stage normalization pipeline that combines LLM-assisted canonicalization with human review. The final pool contains 73 canonical affect concepts, each represented as a structured entry: `(concept_id, canonical_label, definition, strict_aliases, loose_neighbors, valence, arousal, type)`. Figure 5 provides a compact visualization of the semantic organization of the pool. The figure is intended to show the high-level structure of the concept space rather than to enumerate every concept. Table 6 lists the complete set of 73 canonical concepts grouped by their primary high-level affect type.

In the first stage, we perform canonicalization and seed concept extraction. The LLM is used only for lightweight canonicalization of the 248 raw EMER labels: it merges labels with the same or nearly the same affective meaning, selects a unique `canonical_label` for each concept unit, and preserves the remaining surface forms as aliases. Human reviewers then refine the LLM-produced clusters by removing labels that are non-affective, overly coarse, or not reliably supportable from multimodal evidence. The retained seed concepts are assigned concise human-written definitions, yielding 33 seed concepts for semantic expansion.

In the second stage, we perform semantic expansion and concept consolidation. The seed list is treated as an intermediate concept inventory rather than the final pool. The seed concepts are expanded by an LLM into a larger candidate pool of affect concepts, with each candidate containing a canonical label, aliases, and a short definition. The expanded concepts are then reviewed and filtered by humans. We remove concepts that are duplicate, non-affective, overly context-bound, too broad, or difficult to ground from linguistic, acoustic, or visual evidence. After consolidation, the final meta affect concept pool contains 73 canonical concepts.

Table 7 gives representative examples of how raw labels are transformed through Stage I canonicalization and Stage II consolidation. These examples are illustrative rather than exhaustive. Table 8 summarizes common categories of labels removed during consolidation.

Human review is used in both stages to reduce semantic noise and prevent the concept pool from becoming a simple expansion of raw labels. In seed concept extraction, reviewers first remove non-affective labels and then merge obvious synonym clusters. The review pays particular attention to high-overlap regions such as sadness, anxiety/stress, anger, dissatisfaction/disappointment, longing/anticipation, pain/hurt, affiliation/care, and admiration.

We recruited 30 domain-expert reviewers through online academic and professional communities to support concept verification and consolidation. Reviewers had self-reported experience in NLP, affective computing, psychology, linguistics, multimodal analysis, or emotion annotation. They were provided with written guidelines and representative examples before annotation. We collected only minimal demographic information needed for responsible reporting, including self-reported expertise and country or region of residence, and did not collect sensitive personal attributes.

Participants were informed of the research purpose, expected workload, voluntary nature of participation, and data-use policy before beginning the review. They could withdraw at any time. Participants were compensated at approximately US$15 per hour. The compensation was set according to the estimated time required for the task and local norms for expert annotation.

In concept consolidation, borderline candidates are judged according to four criteria. *Groundability* asks whether the concept can be reliably inferred from text, behavior, facial expression, tone, or context. *Transferability* asks whether the concept can

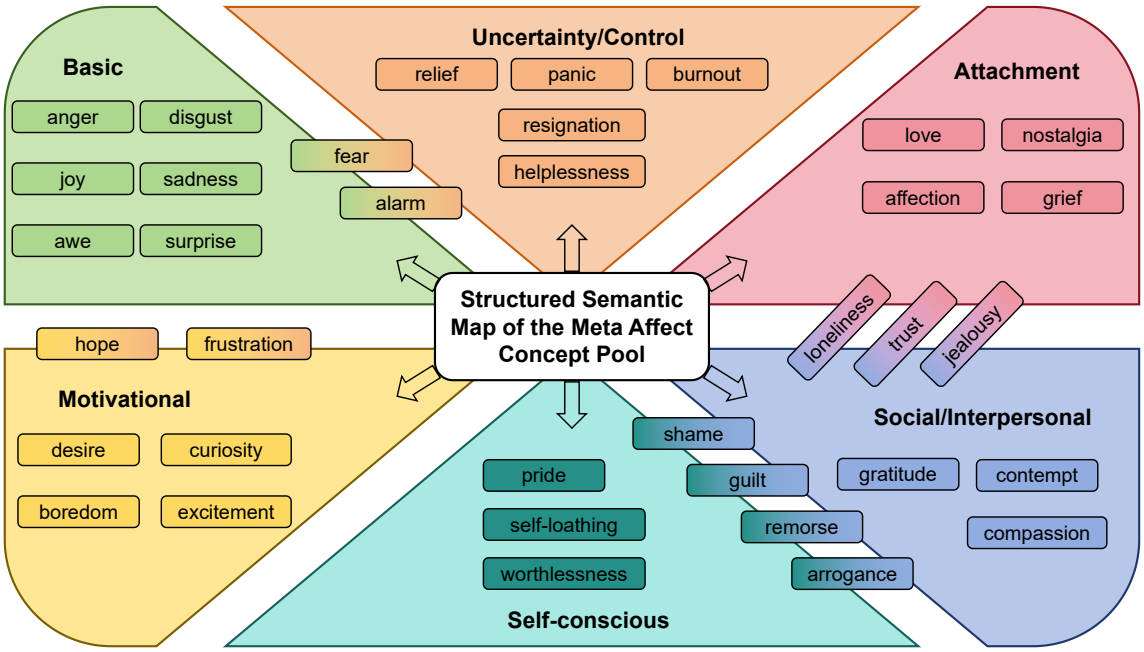

Figure 5: Structured semantic map of the meta affect concept pool. The map visualizes the high-level organization of the pool into six affect types and shows representative concepts from each region. The complete inventory of all 73 canonical concepts is provided in Table 6.

be reused across tasks rather than depending on a specific event, relation, or dataset convention. *Atomicity* asks whether the concept expresses a single affective state rather than mixing intensity, cause, outcome, attitude, or behavior. *Distinctiveness* asks whether the concept has a stable and necessary boundary from neighboring concepts. Only concepts that satisfy these criteria are retained in the final pool.

Finally, each meta affect concept is serialized into a natural-language reranker document by concatenating its canonical label, definition, strict aliases, loose neighbors, valence/arousal attributes, and high-level type. The template is: *"The candidate concept is {canonical_label}. It refers to {definition}. It may also be described as {strict_aliases}. Nearby related concepts include {loose_neighbors}. It is characterized by {valence_text} valence and {arousal_text} arousal, and belongs to {concept_types}."* Unavailable fields are omitted rather than replaced by placeholders.

B Details of Transfer Protocol

This appendix details the zero-shot cross-taxonomy transfer protocol used for MER-UniBench evaluation.

Zero-shot Cross-taxonomy Transfer Protocol. AffectAnchor is trained only with reranked MER-caption+ supervision in the meta affect concept space. During MER-UniBench evaluation, no labeled examples, gold annotations, validation labels, task-specific supervision, prompt tuning, model selection, or hyperparameter selection from MER-UniBench target datasets are used. The target datasets are accessed only through their label names and textual affect descriptions, which serve as inference-time task specifications.

Table 9 summarizes the separation between vocabulary construction, training, and transfer evaluation. EMER labels are used only to construct the fixed meta affect concept pool before MER-UniBench evaluation, MER-caption+ provides training supervision in the meta concept space, and MER-UniBench target resources are used only for inference-time label descriptions and final evaluation. MER-UniBench validation labels, model predictions, and evaluation feedback are not used for reward construction, prompt tuning, hyperparameter selection, or concept-pool revision.

Overlap and provenance checks. Table 9 summarizes the separation among vocabulary construction, training, and evaluation. EMER labels are used only to construct the fixed meta affect concept pool, MER-caption+ is used only for GRPO train-

High-level type	Canonical concepts	Count
Basic	alarm, amusement, anger, awe, calmness, contentment, disgust, distress, fear, joy, sadness, surprise	12
Uncertainty/Control	acceptance, anxiety, burnout, despair, desperation, dread, helplessness, hopelessness, panic, powerlessness, relief, resignation, security, stress	14
Attachment	affection, desolation, emptiness, grief, infatuation, jealousy, loneliness, longing, love, nostalgia, tenderness, trust, yearning	13
Motivational	anticipation, apathy, boredom, curiosity, desire, disappointment, discouragement, dissatisfaction, excitement, frustration, hope	11
Self-conscious	arrogance, confidence, embarrassment, guilt, humiliation, insecurity, pride, remorse, self-loathing, shame, worthlessness	11
Social/Interpersonal	compassion, contempt, distrust, envy, gratitude, hate, hurt, indignation, outrage, paranoia, resentment, suspicion	12

Table 6: Complete inventory of the 73 canonical concepts in the meta affect concept pool, grouped by primary high-level affect type. Concepts with multiple type annotations are listed under one primary type for readability, while their full type annotations are retained in the structured concept entries used by the reranker.

Raw label	Stage-I	Stage-II	Note
happy	delighted	joy	Merged with other high-valence positive labels.
joyful	delighted	joy	Lexical variant consolidated under the same concept.
excited	delighted	excitement	Separated from generic joy due to high arousal.
relaxed	calm	calmness	Direct normalization of low-arousal positive ease.
worried	anxiety	anxiety	Direct uncertainty-related normalization.
terror	fear	dread	Refined into a higher-intensity threat-focused concept.
pressure	overwhelmed	stress	Consolidated as overload- or tension-driven affect.
disappointed	disappointment	disappointment	Retained as an expectation-violation concept.
heartbroken	grief	grief	Retained as an attachment- and loss-related concept.
pessimism	pessimism	hopelessness	Mapped to a stable negative expectancy concept.
caring	warmth	compassion	Interpreted as other-oriented concern.
warm	warmth	affection	Interpreted as affiliative closeness or attachment.
regret	regret	remorse	Normalized to a self-evaluative moral regret concept.
relieved	comforted	relief	Resolved into a post-tension release concept.
shock	shocked	surprise	Direct normalization of abrupt unexpectedness.

Table 7: Representative normalization examples.

Deleted type	Examples	Reason
Polarity-only or overly coarse labels	positive, negative, positive emotion, negativity	Too coarse to serve as evidence-groundable affect concepts.
Dialogue acts / communicative functions	advice, suggestion, inquiry, apology, persuasion, agreement	Conversational functions rather than affective states.
Event/action descriptors	collapse, breakdown, crying, quarrel, arguing, complaint	Events, behaviors, or outcomes rather than affect concepts.
Cognitive or epistemic states	thinking, contemplation, reflection, guessing, understanding	Primarily cognitive rather than affective.
Underspecified intensity or stance markers	serious, intense, severe, strong, urgent, low	Do not specify a distinct affective meaning.

Table 8: Representative deleted label types.

ing and reward references, and MER-UniBench resources are used only for inference-time label descriptions and final evaluation. No MER-UniBench instances, annotations, validation labels, predictions, or evaluation feedback are used for training, reward construction, prompt tuning, hyperparameter/model selection, or concept-pool revision.

We audit possible leakage between MER-caption+ training data and MER-UniBench evaluation splits following the MER-UniBench/AffectGPT protocol. Table 10 lists the checked dimensions. We treat overlap in affect labels or label descriptions as taxonomy overlap rather than instance leakage, since target descriptions are allowed at inference time. In contrast, overlap in samples, media sources, subtitles, descriptions, or speaker/session identifiers would indicate target-instance leakage. For fields with available identifiers or metadata, we find no overlap in sample/video identifiers, source metadata, normalized subtitles, description n-grams, or speaker/session identifiers. For fields without comparable metadata, we rely on official dataset splits and report them as unavailable rather than verified absent.

At inference time, the evidence generator, reranker, and 73-concept meta affect pool are fixed. For each target dataset, only the retrieval candi-

Resource	Role in the transfer protocol
EMER label vocabulary	Used only before evaluation to construct the fixed meta affect concept pool. No MER-UniBench instances or annotations are used.
MER-caption+	Used for GRPO training and reward references in the meta concept space. No target-task supervision is used.
MER-UniBench labels	Used only at inference time as target label descriptions.
MER-UniBench test sets	Used only for final evaluation. Held-out instances and annotations are not used for training, tuning, or pool revision.

Table 9: Data and taxonomy separation in the zero-shot transfer protocol.

Dimension	Compared fields	Treatment
Sample identity	sample IDs, video IDs	exact match when available
Media source	source IDs, URLs, dataset metadata	provenance match when available
Subtitles	normalized transcript strings	exact and high-similarity text match
Descriptions	normalized affect-description n-grams	near-duplicate content check
Speaker/session	speaker IDs, session IDs, dialogue IDs	train–test reuse check when available
Label vocabulary	affect labels and label descriptions	reported as taxonomy overlap, not instance leakage

Table 10: Overlap audit between MER-caption+ training resources and MER-UniBench evaluation splits. Identifier- and metadata-based checks are applied when the corresponding fields are available.

date set is replaced by target-side affect descriptions. Given a multimodal input, AffectAnchor generates affective evidence and ranks candidate descriptions with the frozen reranker; predictions are then obtained according to the corresponding MER-UniBench protocol.

For sentiment datasets, the candidate pool contains positive, negative, and neutral/mixed polarity descriptions. For basic emotion datasets, it contains the dataset-specific emotion taxonomy, with each category represented by a short affective description. For OV-MERD+, we follow the AffectGPT/MER-UniBench evaluator, which normalizes predicted and reference labels through the emotion-wheel mapping before computing grouped open-vocabulary scores.

Under this protocol, MER-UniBench evaluates zero-shot cross-taxonomy transfer rather than supervised target-task adaptation. The zero-shot claim is restricted to target-task instances, annotations, validation labels, and task-specific supervision; it does not imply the absence of external affect knowledge, since the intermediate concept pool is constructed from EMER labels.

Task-specific Evaluation Metrics. The evaluation metrics follow the official MER-UniBench protocol and differ across task categories. For sentiment analysis datasets, including CMU-MOSI, CMU-MOSEI, CH-SIMS, and CH-SIMS v2, we report Weighted Average F-score (WAF). WAF is used because sentiment labels are often imbalanced. It computes the F-score for each class and averages them according to class support, so that both frequent and infrequent sentiment categories contribute to the final score.

For basic emotion recognition datasets, including MER2023, MER2024, MELD, and IEMOCAP, we report Hit Rate (HIT). HIT is designed for the multi-label setting used by MER-UniBench, where a sample may be associated with multiple plausible basic emotion labels. A prediction is counted as a hit if the predicted candidate set contains at least one reference emotion label according to the benchmark protocol. This metric is therefore better aligned with multi-label basic emotion recognition than strict exact-match accuracy.

For fine-grained and open-vocabulary emotion recognition, including OV-MERD+, we report grouped F_S scores following the official evaluator. Open-vocabulary affect labels may differ in surface form while expressing similar affective meanings. The evaluator therefore normalizes predicted and reference labels through the emotion-wheel mapping and computes scores over grouped affect categories. This reduces the penalty for harmless synonym variation while still evaluating whether predictions fall into the correct fine-grained affect region.

The overall MER-UniBench score reported in the main paper is the dataset-wise mean of these official primary metrics. We use this average only as a compact summary across heterogeneous benchmarks; per-dataset scores are also reported to show performance under each task-specific metric.

C Implementation Details

This appendix supplements the implementation details. All main AffectAnchor results are reported from a single GRPO training run and the final evaluated checkpoint. We do not select checkpoints

Item	Value
Backbone	Qwen2.5-Omni-7B
Training method	LoRA + GRPO
LoRA rank / alpha / dropout	32 / 64 / 0.05
LoRA target modules	all-linear
Frozen module	visual encoder
Quantization	none
Reranker	Qwen3-Reranker-8B
Concept pool size	73
Top-k	5
Reward	$R_{F1} + R_{\mathrm{N@5}}$
Reward weights	$\lambda_F = 1.0, \lambda_N = 1.0$
Group size G	4
Learning rate	1×10^{-5}
Per-device batch size	4
Gradient accumulation	1
Processes per node	4
Effective completion batch	16
Training epochs	1
Optimizer	adamw_torch_fused
Adam $\beta_1/\beta_2/\epsilon$	0.9 / 0.95 / 10^{-8}
Weight decay	0.1
Warmup / scheduler	0.01 / cosine
KL coefficient	0
GRPO clip range	$[0.8, 1.2]$
Loss type	dr_grpo

Table 11: Main implementation and GRPO hyperparameters used for AffectAnchor.

using MER-UniBench validation labels or target-task feedback. This follows the reporting style of recent MER-UniBench baselines, which typically provide single-score benchmark results rather than multi-seed confidence intervals. Accordingly, our SOTA comparison should be interpreted as a point-estimate comparison under the official evaluation protocol. We leave multi-seed training variance and paired significance testing with released prediction files to future work. The main training and optimization hyperparameters are summarized in Table 11.

Dataset and Resource Statistics. Table 12 summarizes the resources used for meta concept pool construction, GRPO training, and zero-shot evaluation. The meta affect concept pool is initialized from the label vocabulary of EMER, which is used only as a label source for constructing the intermediate affect-semantic space. Starting from the raw EMER affect labels, we obtain 33 seed concepts after canonicalization and human review, and 73 final canonical concepts after semantic expansion and consolidation.

For policy optimization, we use MER-caption+, which contains approximately 31K multimodal samples with video, audio, subtitles, emotion descriptions, and open-vocabulary affect labels. MER-UniBench target datasets are used only for final zero-shot evaluation. No MER-UniBench training or validation labels are used for model training, reward construction, prompt tuning, hyperparameter selection, or concept pool revision.

Artifact licenses and intended use. We use existing datasets, benchmarks, and pretrained models according to their original licenses, access conditions, and intended research use. EMER labels are used only for meta concept pool initialization, MER-caption+ is used for GRPO training supervision, and MER-UniBench target datasets are used only for final zero-shot evaluation. We do not redistribute original dataset instances. Any released derived artifacts, including the meta affect concept pool, prompts, evaluation scripts, or trained adapters, should be accompanied by documentation of source artifacts, applicable license requirements, and intended research-use boundaries. The artifacts created in this work are intended for research on evidence-grounded multimodal affect understanding and should not be used as standalone tools for high-stakes decision-making, surveillance, diagnosis, hiring, education assessment, or law enforcement.

Model and Reranker. AffectAnchor is initialized from Qwen2.5-Omni-7B and trained with LoRA while freezing the visual encoder. We use LoRA rather than QLoRA, without quantized loading. The LoRA rank, alpha, and dropout are 32, 64, and 0.05, respectively, and LoRA adapters are applied to all linear modules.

The meta affect concept pool contains 73 canonical concepts and is fixed across training and evaluation. Each concept is serialized as a natural-language reranker document containing its canonical label, definition, strict aliases, loose neighboring concepts, valence/arousal attributes, and high-level type. Unavailable fields are omitted. The same serialization is used whenever the reranker scores generated evidence against concept descriptions. Details of the pool construction and serialization template are provided in Appendix A.

We use Qwen3-Reranker-8B as a frozen reranker. During reward computation, the query contains only model-generated affective evidence, without gold label hints or reference affect labels. Candidate concepts are drawn from the fixed 73-concept pool, and gold relevant concepts are taken from the

Resource	Role	Size / Split	Supervision / Access
EMER label vocabulary	Concept pool initialization	248 raw labels	Label vocabulary only
Meta concept pool	Retrieval target space	73 concepts	Concept-level reward space
MER-caption+	GRPO training	∼31K samples	Training reward references
CMU-MOSI	Zero-shot evaluation	Official test split, 686 samples	Final evaluation only
CMU-MOSEI	Zero-shot evaluation	Official test split, 4,659 samples	Final evaluation only
CH-SIMS	Zero-shot evaluation	Official test split, 457 samples	Final evaluation only
CH-SIMS v2	Zero-shot evaluation	Official test split, 1,034 samples	Final evaluation only
MER2023	Zero-shot evaluation	Official test split, 411 samples	Final evaluation only
MER2024	Zero-shot evaluation	Official test split, 1,169 samples	Final evaluation only
MELD	Zero-shot evaluation	Official test split, 2,610 samples	Final evaluation only
IEMOCAP	Zero-shot evaluation	Official test split, 1,241 samples	Final evaluation only
OV-MERD+	Zero-shot evaluation	Official test split, 532 samples	Final evaluation only

Table 12: Dataset and resource statistics. EMER labels are used only to initialize the meta affect concept pool. MER-caption+ provides training reward references in the meta concept space. Target benchmark instances and annotations are used only for final evaluation, not for training, reward construction, prompt tuning, hyperparameter selection, or concept-pool revision.

top-5 ranked targets in the reranked MER-caption+ annotations. In all main experiments, we use $k = 5$ and set the reward weights to $\lambda_{\text{F1}} = \lambda_{\text{N}} = 1.0$, yielding $R_{\text{ret}} = R_{F1} + R_{\text{N@5}}$.

GRPO Training. No supervised fine-tuning stage is used before GRPO. Starting from the original Qwen2.5-Omni-7B checkpoint, we directly optimize LoRA adapters with the retrieval reward. AffectAnchor is optimized with GRPO for one epoch, without a fixed maximum number of steps. For each prompt, we sample $G = 4$ completions. Rewards are normalized within the completion group from the same prompt, following the default group-level reward scaling. The per-device training batch size is 4, gradient accumulation is 1, and training uses 4 processes per node, giving a completion-level effective batch size of 16 and approximately 4 unique prompts per optimization step.

The optimizer follows the ms-swift/HuggingFace default used in the saved GRPO configuration, namely `adamw_torch_fused`. We use learning rate 1×10^{-5}, warmup ratio 0.01, cosine scheduling, and weight decay 0.1. Although the general GRPO objective may include KL regularization, our main run sets both `BETA` and `KL_COEF` to 0, so no active KL penalty or separate reference model is used. The GRPO clipping range is $[0.8, 1.2]$, and the loss type follows the default `dr_grpo`. The reranker batch size is 8, but this only controls reward HTTP request chunking and is not the training batch size.

Baselines and Resources. For fair comparison, all baselines use the same input modalities, prompts, decoding settings, and evaluation protocols whenever supported. In MER-UniBench evaluation, no labeled examples or gold annotations from target datasets are used for training, reward construction, prompt tuning, or hyperparameter selection; the target-side affect description pool is replaced according to each dataset taxonomy, as detailed in Appendix B. Training is conducted on four NVIDIA A100-80G GPUs and takes approximately two days.

Item	Count
Sampled OV-MERD examples	100
Systems per example	4
Model responses	400
Reference evidence units	1,830
Generated claims	2,992
Annotators	2
Calibration examples	10
Human-adjudicated examples	100

Table 13: Statistics of the human validation subset. The 10 calibration examples are excluded from the reported agreement and metric computation.

Pair	Acc.	F1	κ_w
Human A/B	80.61	77.16	0.814
GPT-5.4/Human	78.11	74.69	0.783
Open Judge/Human	74.93	71.24	0.739

Table 14: Claim-level agreement. Acc. and F1 are percentages; κ_w is linearly weighted Cohen's κ.

Human Validation and Judge Sensitivity Our evidence-level evaluation is LLM-assisted: GPT-5.4 is used to decompose descriptions and model responses into atomic units and to align generated claims with hidden reference evidence units. To assess the reliability of this evaluation protocol,

Model	P	R_w	Hall.↓	Cov.
E-LLaMA	52.84	43.92	40.31	49.85
AffectGPT	52.10	49.28	44.05	51.44
Qwen-Omni	60.96	60.38	32.06	66.52
AffectAnchor	65.88	72.91	29.34	78.66

Table 15: Human-adjudicated evidence scores. P: precision; R_w: weighted recall; Cov.: modality coverage.

Verdict	$v(c)$	Precision	Recall cover
Supported	1.0	full	yes
Partial	0.5	half	yes
Unsupported	0.0	none	no
Contradicted	-1.0	none	no

Table 16: Claim-to-reference alignment verdicts. Unsupported and contradicted claims are counted as hallucinations.

we conduct an additional human validation and judge sensitivity study. This study focuses on the claim-to-reference alignment stage, while the atomized reference units and generated claims are also manually inspected during annotator calibration to remove malformed or non-atomic units.

D Evidence Evaluation Details

This appendix provides additional details for the zero-shot evidence evaluation on OV-MERD. The goal is to measure whether a model recovers the affect-relevant evidence underlying the ground-truth description, rather than only predicting the final affect label.

Atomic Evidence Extraction. We evaluate all models on the same OV-MERD samples. For each sample, we use GPT-5.4 to decompose the OV-MERD description into hidden reference evidence units. Each reference unit contains a unit identifier, modality, temporal phase, evidence type, atomic content, and source span. The modality is one of text, audio, visual, or integration (cross-modal). The evidence type is either observable or inferential. Observable units correspond to directly described cues, such as subtitles, facial expressions, gestures, vocal tone, or pauses. Inferential units correspond to affective interpretations supported by these cues, such as frustration, concern, teasing, or uncertainty.

For each model response, we also use GPT-5.4 to decompose the generated reasoning response into atomic claims. Each generated claim contains a claim identifier, claim text, source span, modality focus, temporal phase, and claim level. The claim level distinguishes local evidence, integrated inference, final affect summary, and other auxiliary statements. This produces two comparable atomized sources for each sample: hidden reference units extracted from the OV-MERD description and generated claims extracted from the model response.

Claim-to-Reference Alignment. We align each generated claim with the hidden reference units of the same sample. For a generated claim c, the alignment verdict is one of four categories: supported, partially supported, unsupported, or contradicted. We map these verdicts to scalar values as shown in Table 16.

For claim-level scoring, if a generated claim can be aligned with multiple reference units, we use its best supported alignment. For reference-side recall, a reference unit is counted as covered when at least one generated claim is aligned to it as supported or partially supported. Unsupported and contradicted claims do not contribute to reference coverage.

Metric Computation. Let N_c denote the number of generated claims for a sample. Evidence precision measures the supported portion of generated claims:

$$\text{Prec.} = \frac{1}{N_c} \sum_{c=1}^{N_c} \max(v(c), 0). \tag{10}$$

This is equivalent to

$$\frac{N_{\text{supported}} + 0.5 N_{\text{partially supported}}}{N_c}, \tag{11}$$

because contradicted claims are clipped to zero by $\max(v(c), 0)$.

Hallucination rate measures the proportion of unsupported or contradicted generated claims:

$$\text{Halluc.} = \frac{N_{\text{unsup}} + N_{\text{contra}}}{N_c}. \tag{12}$$

where N_{unsup} and N_{contra} denote the numbers of unsupported and contradicted claims. Partially supported claims are not counted as hallucinations, but receive only half credit in evidence precision.

Weighted evidence recall measures how much important reference evidence is recovered:

$$\text{Wtd.Recall} = \frac{\sum_{e \in E_{\text{covered}}} w(e)}{\sum_{e \in E_{\text{all}}} w(e)}. \tag{13}$$

Here, E_{all} is the set of all hidden reference units for a sample, and E_{covered} is the subset covered by

Task	Dataset	Metric	AffectGPT	AffectAnchor
Sentiment	CMU-MOSI	ACC	81.25	85.37
		WAF	**81.30**	**85.44**
	CMU-MOSEI	ACC	80.68	85.91
		WAF	**80.90**	**85.95**
	CH-SIMS	ACC	88.40	88.14
		WAF	**88.49**	**88.06**
	CH-SIMS v2	ACC	86.17	87.82
		WAF	**86.18**	**87.78**
Basic	MER2023	**HIT**	**78.54**	**89.70**
	MER2024	**HIT**	**78.80**	**91.12**
	MELD	**HIT**	**55.65**	**68.68**
	IEMOCAP	**HIT**	**60.54**	**75.26**
Fine-grained	OV-MERD+	F_S	**62.52**	**66.71**
		$Precision_S$	62.21	59.89
		$Recall_S$	63.00	75.68
Overall	Avg.	Primary mean	74.77	82.08

Table 17: Additional MER-UniBench metrics comparing AffectAnchor with AffectGPT. Official primary metrics are shown in bold: WAF for sentiment datasets, HIT for basic emotion datasets, and F_S for open-vocabulary emotion recognition.

at least one supported or partially supported generated claim. We assign weight 1.0 to observable evidence and 0.5 to inferential evidence, giving higher importance to directly grounded multimodal cues while still crediting recovered affective interpretations.

Modality coverage measures whether generated claims cover the evidence-bearing modalities present in the reference units. We consider four modality groups: text, audio, visual, and integration. A modality is counted as covered if at least one generated claim with the corresponding modality focus is aligned to a reference unit from that modality as supported or partially supported. The final modality coverage score is computed at the sample level and then averaged across samples.

We sample 100 OV-MERD examples from the evidence evaluation set, stratified by affect category and evidence-bearing modality. For each example, we include responses from four systems: E-LLaMA, AffectGPT, Qwen-Omni, and AffectAnchor. System identities are anonymized before annotation. The validation subset contains 400 model responses, 1,830 reference evidence units, and 2,992 generated claims. We use 10 additional examples for annotator calibration; these calibration examples are excluded from all reported agreement and metric computations. Table 13 summarizes the subset statistics.

We recruited two trained annotators from the same online expert-reviewer pool described in Appendix A for the human validation study. Both annotators had self-reported experience in NLP, affective computing, multimodal analysis, or emotion annotation. They were compensated at US$15 per hour. Before annotation, they were informed of the research purpose, expected workload, voluntary nature of participation, compensation policy, and data-use policy, and they could withdraw at any time.

The two annotators independently judged each generated claim against the hidden reference evidence units from the same example. For each claim, annotators assigned one of four verdicts: *supported*, *partially supported*, *unsupported*, or *contradicted*. The annotators were instructed to judge only whether the claim was supported by the provided reference evidence, without considering the source system or the final emotion prediction. Disagreements were resolved through adjudication, and the adjudicated labels were used as the human reference for automatic-judge comparison.

We report exact agreement, macro-F1, and weighted Cohen's κ over the four ordinal verdict labels. Weighted κ uses linear ordinal weights, treating disagreements between adjacent labels as less severe than disagreements between *supported* and *contradicted* claims. As shown in Table 14, the

Prompt family	Input	Output	Key constraints
Stage I canonicalization	Raw OV-MER emotion labels	Canonical concept units with `canonical_label` and `aliases`	Preserve every input label exactly once; merge only near-equivalent labels; keep uncertain cases separate.
Stage II semantic expansion	Human-reviewed seed concepts	Expanded candidate concepts with labels, aliases, and definitions	Generate affective concepts only; remove cognitive, behavioral, polarity-only, or stylistic descriptors; return structured JSON.
Concept scoring and consolidation	Candidate concept pool	Scores, structural attributes, keep/drop recommendations, and merge/split suggestions	Score groundability, transferability, atomicity, and distinctiveness; LLM outputs are used only as decision support for human review.
Description-to-evidence extraction	OV-MERD descriptions	Hidden atomic reference evidence units	Split descriptions into minimal multimodal evidence units; preserve source spans; label modality, phase, and evidence type; do not add information absent from the description.
Response-to-claim extraction	Model-generated reasoning responses and reference-unit file for ID checking	Model-generated atomic claims	Extract claims only from the response; use reference units only for sample-id alignment and format checking; do not judge, filter, correct, or rewrite claims based on references.
Atomic evidence alignment judge	Hidden reference evidence units and model-generated atomic claims	Claim-level verdicts and sample-level grounding scores	Use only provided reference units; do not invent evidence; assign supported, partially supported, unsupported, or contradicted verdicts.

Table 18: Summary of LLM prompt families used for concept-pool construction and evidence evaluation.

two human annotators achieve substantial agreement. GPT-5.4 also shows substantial agreement with the adjudicated human labels, while the open-weight judge obtains slightly lower but still consistent agreement. These results indicate that the LLM-assisted alignment is reasonably consistent with human evidence judgments.

We recompute evidence precision, hallucination rate, weighted evidence recall, and modality coverage using the adjudicated human labels. Table 15 shows that the human-adjudicated scores preserve the same system-level conclusion as the full GPT-5.4-based evaluation: AffectAnchor improves evidence recall and modality coverage over the Qwen-Omni backbone and affect-oriented baselines, while maintaining a lower hallucination rate. The absolute values differ from the full automatic evaluation because the subset is smaller and human judgments are more conservative, but the relative ranking remains stable.

To test whether the conclusion depends on GPT-5.4 specifically, we rerun the claim-to-reference alignment stage with an open-weight instruction model under the same prompt format and deterministic decoding. The open judge uses the same reference units and generated claims, and system identities remain anonymized. Although the open judge produces slightly lower agreement with human labels than GPT-5.4, it preserves the same qualitative ranking among systems. This indicates that the evidence-evaluation conclusion is not specific to a single proprietary judge.

In Tables 14 and 15, Acc. and F1 are percentages, κ_w denotes linearly weighted Cohen's κ, P denotes evidence precision, R_w denotes weighted evidence recall, Hall. denotes hallucination rate, and Cov. denotes modality coverage.

Aggregation. All four metrics are first computed for each sample and then averaged across the evaluation set. For multi-model comparisons, we use the same hidden reference units for all models, so metric differences reflect differences in generated reasoning responses rather than differences in reference decomposition. Prompt templates for description decomposition, response decomposition, and claim-to-reference alignment are provided in Appendix F.

E Complete MER-UniBench Results

Table 17 reports the complete results, with several metrics for each dataset, and the primary metrics are shown in bold. Because several recent baselines report only the official MER-UniBench primary scores and do not release complete auxiliary metrics or prediction files, we report additional metrics only against AffectGPT. This auxiliary comparison complements the main SOTA comparison in Table 2 and is not used to replace the official primary-metric ranking.

F LLM Prompts

This section summarizes the prompt templates used for meta affect concept pool construction and zero-shot evidence evaluation. We use six prompt families. The first three prompts support construction of the meta affect concept pool: concept canonicalization, semantic expansion, and concept consolidation. In these prompts, LLM outputs are treated as candidates or decision aids and are manually inspected before being included in the final pool.

The remaining three prompts support evidence evaluation. Description-to-evidence extraction decomposes OV-MERD descriptions into hidden atomic reference evidence units. Response-to-claim extraction decomposes model-generated reasoning responses into atomic claims, using the reference-unit file only for sample-id alignment and format checking. Atomic evidence alignment then compares generated claims with hidden reference units and assigns claim-level verdicts.

Table 18 lists the input, output, and key constraints of each prompt family. Across all prompts, we emphasize structured outputs, explicit decision criteria, and conservative handling of ambiguous cases. For concept-pool construction, the prompts require coverage checks, JSON-formatted outputs, and separation between strict aliases, loose semantic neighbors, and conflicting aliases. For evidence evaluation, the prompts separate extraction from judging: GPT-5.4 first atomizes descriptions and responses, and the later alignment prompt is restricted to the supplied reference evidence units.

We provide summaries here to keep the appendix concise, while preserving the information needed to understand the role and constraints of each prompt family.