CONTEXTUAL CUE REPLAY: TRAINING-FREE WHOLE-IMAGE EMOTION RECOGNITION VIA PROGRESSIVE RE-READING IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have made substantial progress in visual understanding, yet whole-image emotion recognition remains challenging, as it requires interpreting visual cues in context and weighing their contributions to the overall emotional impression. Emotion-specific fine-tuning incurs training and data-preparation costs when adapting MLLMs to different emotion-category systems, motivating training-free approaches. While existing training-free methods such as visual token selection help retain relevant evidence, such filtering alone does not establish an explicit feedback path for task-conditioned scene information to refine visual representations. In image-first autoregressive MLLMs, causal attention prevents visual positions from directly accessing such information formed at subsequent instruction positions. We propose **Contextual Cue Replay (CCR)**, a training-free framework that enables progressive contextual re-reading within frozen MLLMs. At selected decoder depths, an auxiliary block evaluation constructs contextualized visual states with access to the complete image and image-conditioned states of a generic task instruction. The base and contextualized states then query a shared memory of the current visual representations. Their retrieval difference captures context-induced changes in retrieved evidence and is written back as an incremental correction to the visual residual. Native decoder blocks integrate each correction before the next replay stage, allowing subsequent re-reading to build on visual evidence and instruction states shaped by earlier updates. Experiments across diverse benchmarks and emotion-category systems demonstrate that CCR yields consistent performance improvements without parameter updates or image-specific human guidance.

1 INTRODUCTION

Multimodal large language models (MLLMs) have advanced in recognizing, describing, and reasoning about visual content (Bai et al., 2025). Yet recognizing what is in an image is not enough to understand how the image feels as a whole (She et al., 2026; Wu et al., 2026). Whole-image emotion recognition requires interpreting local emotional cues in relation to the broader scene (Lee et al., 2026), considering how subjects' states, scene relationships, and the atmosphere created by color, lighting, and composition jointly shape an overall emotional impression (Sadeghi et al., 2024).

Images often contain multiple, interrelated cues whose emotional meanings and contributions depend on context. This dependence presents two distinct challenges: meaning ambiguity, where a local cue admits different emotional in-

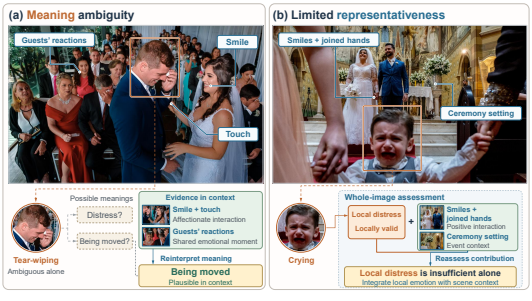


Figure 1: Context shapes emotional cue meaning and contribution. (a) The bride's and guests' reactions support being moved as a plausible interpretation of the groom's tear-wiping. (b) The child's crying still signals local distress, but the couple's smiles and ceremony context call for reassessing its contribution to whole-image emotion.

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

terpretations, and limited representativeness, where a cue with a clear local meaning may not adequately represent the image as a whole (Fig. 1). Effective recognition therefore requires identifying relevant cues, interpreting their meanings in context, and weighing their contributions to a whole-image judgment.

Existing efforts to enhance emotion understanding in MLLMs have explored emotion-specific fine-tuning (Xie et al., 2024; Yang et al., 2024) and visual guidance (Zhang et al., 2024). Fine-tuning incurs training and data-preparation costs when adapting models to different emotion-category systems, while manually specified regions or explanations add image-specific effort. These requirements motivate training-free approaches that support flexible reuse of frozen models. SEPM (Fang et al., 2025) advances this direction through confidence-guided coarse-to-fine classification and emotion-oriented visual token selection. Its visual intervention determines which evidence is retained for subsequent inference. However, selecting relevant content is distinct from revising how that content is represented and interpreted in light of the overall scene. We therefore investigate how task-conditioned scene information can be explicitly fed back into visual representations, allowing context to guide their progressive refinement during inference.

An intuitive approach is **progressive re-reading**: take in the scene as a whole, revisit its details in light of that understanding, and use the reinterpreted evidence to refine the overall judgment. In autoregressive MLLMs where image tokens precede the task instruction, however, causal attention allows instruction states to incorporate visual information but prevents visual positions from consulting those subsequent states (Pei et al., 2025). Task-conditioned scene information formed at instruction positions therefore lacks a direct route back to visual representations, motivating an explicit feedback mechanism during inference.

We propose **Contextual Cue Replay (CCR)**, a training-free framework that implements this feedback within a frozen MLLM. At selected decoder depths, an auxiliary contextual evaluation allows visual positions to access the complete image and image-conditioned states of a generic task instruction. The current visual states and their contextualized counterparts then query a shared memory of the current visual representations. Their retrieval difference captures how contextualization changes the retrieved evidence and is written back as an incremental correction to the visual residual. Subsequent native decoder blocks integrate this correction into visual and instruction representations, allowing the next replay stage to build on evidence and context shaped by earlier updates. Experiments across diverse benchmarks demonstrate consistent performance improvements without additional training or image-specific human guidance.

The main contributions of this work are as follows:

- **Contextual feedback perspective.** We introduce progressive contextual re-reading for whole-image emotion recognition, using task-conditioned scene understanding to guide visual evidence refinement.
- **Training-free framework.** CCR combines auxiliary contextualization, shared-memory differential retrieval, and progressive residual write-back within frozen MLLMs.
- **Empirical validation.** Consistent gains across diverse benchmarks and emotion-category systems validate the effectiveness of inference-stage visual evidence refinement.

2 Related Work

Visual Emotion Understanding with MLLMs. One line of MLLM-based research develops visual emotion understanding through emotion-oriented instruction tuning and specialized visual representations (Cheng et al., 2024). EmoVIT (Xie et al., 2024) uses GPT-assisted emotion instruction data for model adaptation, while EmoLLM (Yang et al., 2024) combines content- and relation-based perspectives through multi-perspective visual projection. Such approaches involve emotion-specific supervision and parameter adaptation, introducing additional training and data-preparation costs. CCR uses task-conditioned scene information at inference time to progressively refine visual evidence within a fully frozen model.

Training-Free Inference Enhancement in MLLMs. Training-free methods enhance MLLMs by modifying prompts, inference procedures (Leng et al., 2024), or visual processing without updating model parameters (Liu et al., 2025). Prompt-based and multi-step approaches structure the

inference process (Wei et al., 2022; Kojima et al., 2022), while visual token sparsification methods, such as SparseVLM (Zhang et al., 2025), use textual guidance to reduce visual redundancy and improve computational efficiency. Related work on future-aware attention investigates relaxing causal masking for visual queries, enabling access to subsequent visual or textual context (Pei et al., 2025). Within emotion recognition, SEPM (Fang et al., 2025) combines confidence-guided coarse-to-fine inference with Focus-on-Emotion visual augmentation. It adjusts fine-grained classification prompts using coarse-stage predictions and confidence, and uses attention from emotion-focused prompt tokens to select the visual tokens retained for subsequent inference. CCR shifts the focus from selecting which evidence to retain to refining how that evidence is represented in context, while preserving the complete visual-token sequence. Building on future-aware attention, CCR uses auxiliary contextualized states and their main-branch counterparts to query a shared visual memory, writing the retrieval difference back as an incremental correction to the visual residual. This intervention is repeated at selected decoder depths, with native blocks propagating each correction so that later stages can use evidence and context shaped by earlier updates.

3 METHOD

In image-first autoregressive MLLMs, causal attention lets instruction states integrate image and task information but prevents visual positions from consulting the subsequent instruction. This asymmetry motivates an explicit pathway for contextual re-reading of cues whose emotional contributions depend on the surrounding scene. As shown in Fig. 2, Contextual Cue Replay (CCR) uses an auxiliary block evaluation to construct visual states conditioned on the complete image and task instruction. These states and their main-branch counterparts query a shared visual memory, and their retrieval difference is written into the visual residual. Native blocks integrate each correction before the next replay stage, so later re-reading builds on earlier updates. The frozen output head produces the whole-image emotion prediction.

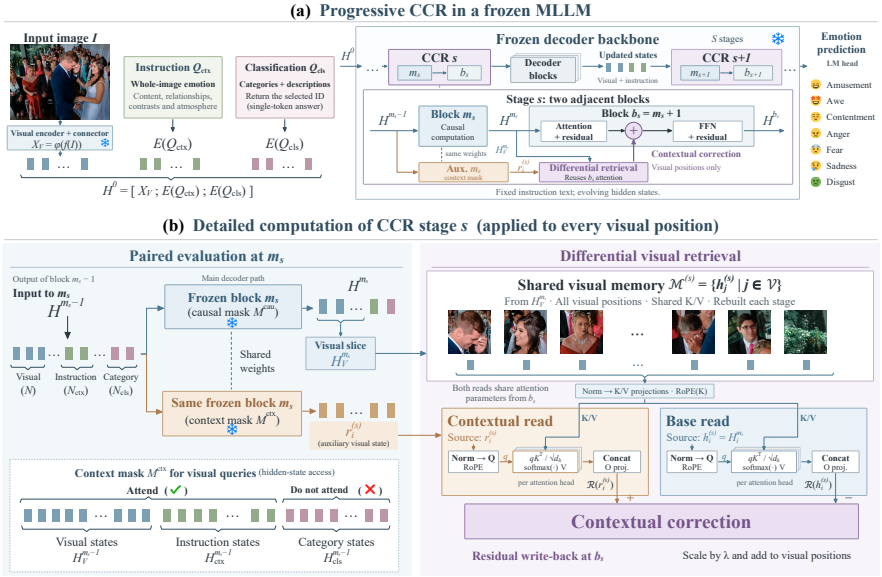


Figure 2: Overview of Contextual Cue Replay (CCR). (a) CCR refines visual representations across selected decoder depths, with native blocks propagating updates between stages. (b) Base and contextualized states query the same complete visual memory. Their retrieval difference is scaled by λ and added to visual post-attention residuals before the feed-forward transformation. All model parameters remain frozen.

3.1 TASK AND INPUT REPRESENTATION

Given an image I and emotion categories $\mathcal{Y} = \{y_1, \ldots, y_K\}$, the objective is to predict the category describing the image as a whole. The frozen backbone comprises a visual encoder f, a visual-to-

language connector ϕ, and a causal language decoder that jointly processes visual and text tokens (Liu et al., 2023). Inference requires only the image and generic task and category prompts.

The encoder and connector produce N visual embeddings $X_V = \phi(f(I)) \in \mathbb{R}^{N\times d}$, where d is the decoder hidden dimension. The decoder input is

$$H^0 = \big[X_V;\, E(Q_{\mathrm{ctx}});\, E(Q_{\mathrm{cls}})\big], \tag{1}$$

where E is the frozen text-embedding lookup. Fixed system prompts and conversation delimiters follow the backbone template and are omitted from Eq. (1).

The contextual instruction Q_{ctx} is shared across images: "Judge the emotional character of the image as a whole. Consider visible content, relationships, contrasts, and the overall color, lighting, and composition."

The classification prompt Q_{cls} lists the categories and their descriptions, assigns each category an identifier, and requests the selected identifier. Each identifier occupies one token at the answer position; a_k denotes the vocabulary index of the identifier for category y_k.

Let $\mathcal{V}$ denote all visual-token positions and $\mathcal{C}$ the prefix ending at Q_{ctx}, containing the complete visual sequence and preceding template tokens but excluding the classification suffix Q_{cls}. The instruction states are image-conditioned through causal attention. The contextual source mask below restricts feedback to $\mathcal{C}$, keeping visual updates independent of the candidate-category menu at every stage. CCR operates on every visual position.

We use a pre-normalized causal decoder (Xiong et al., 2020) with L blocks and rotary positional embeddings (Su et al., 2024). Let $\mathcal{B}_\ell(X; M)$ denote its complete frozen block under mask M, including attention, residual connections, and the feed-forward transformation. The native mask M^{cau} permits keys at $j \le i$ for a query at i, and H^ℓ denotes the main-branch output after block ℓ, including completed CCR updates.

After native processing through block $m_1 - 1$, CCR applies S stages. Stage s uses block m_s for contextualization and $b_s = m_s + 1$ for retrieval and write-back. Native blocks between stages propagate the updated evidence.

3.2 Context- and Task-Conditioned Re-reading

CCR uses an auxiliary block evaluation to let each visual position consult the complete image and image-conditioned instruction states. Paired main and auxiliary evaluations provide base and contextualized query sources at the same decoder depth for the subsequent retrieval comparison.

At stage s, both evaluations start from the current main sequence H^{m_s-1}. The main branch applies block m_s with its causal mask, while the auxiliary evaluation uses the following mask over the source prefix $\mathcal{C}$:

$$M^{\mathrm{ctx}}_{ij} = \begin{cases} 0, & i \in \mathcal{V},\ j \in \mathcal{C}, \\ -\infty, & i \in \mathcal{V},\ j \notin \mathcal{C}, \\ M^{\mathrm{cau}}_{ij}, & i \notin \mathcal{V}. \end{cases} \tag{2}$$

This access pattern follows the principle of future-aware attention for visual queries (Pei et al., 2025). The two evaluations produce

$$H^{m_s} = \mathcal{B}_{m_s}\big(H^{m_s-1}; M^{\mathrm{cau}}\big), \quad r_i^{(s)} = \big[\mathcal{B}_{m_s}\big(H^{m_s-1}; M^{\mathrm{ctx}}\big)\big]_i, \quad i \in \mathcal{V}. \tag{3}$$

We call $h_i^{(s)} = H_i^{m_s}$ the **base state** and $r_i^{(s)}$ the **contextualized state**. The base state continues along the main branch and includes the effects of preceding CCR stages; the contextualized state serves as a temporary query source for block b_s. Both evaluations derive their keys and values from H^{m_s-1}, retain the block's residual and feed-forward transformations, and preserve the original token positions. Only the visual rows require auxiliary computation.

3.3 Differential Visual Retrieval

CCR compares the paired states' retrieval outputs from a shared, complete visual memory to extract context-induced changes in retrieved evidence (Fig. 2(b)). The resulting correction refines the current visual residual rather than replacing it.

At stage s, the memory consists of the base states $h_j^{(s)}$, $j \in \mathcal{V}$. It is held fixed during the two reads and rebuilt from the evolving main branch at each subsequent stage. Let $\mathcal{R}_{s,i}(x)$ denote the projected attention output obtained by reading this memory with query-source state x at position i. The **contextual correction** is

$$\Delta_i^{(s)} = \mathcal{R}_{s,i}\Big(r_i^{(s)}\Big) - \mathcal{R}_{s,i}\Big(h_i^{(s)}\Big). \tag{4}$$

The base read serves as a reference with the same complete visual memory as the contextualized read, including positions later than i. These complete-memory reads are distinct from the native causal attention and share the receiving block's parameters and positional indices. Subtracting the base read focuses the update on what context changes. Simply adding the full contextual read would also add the base retrieval itself. The difference provides an incremental correction that becomes zero when contextualization leaves the retrieved evidence unchanged.

Both reads reuse block b_s's frozen pre-attention normalization, query/key/value projections, head assignments, attention scaling, and output projection O_{b_s}, with queries and keys retaining their original rotary positions. The projected read is

$$\mathcal{R}_{s,i}(x) = O_{b_s}\left(\mathrm{Concat}_a\left[\sum_{j\in\mathcal{V}} \alpha_{s,ij}^a(x) v_{s,j}^a\right]\right). \tag{5}$$

For head a, $\alpha_{s,ij}^a(x)$ is scaled dot-product attention, softmax-normalized over all visual positions. Its query is formed from x, while its keys and values $v_{s,j}^a$ are formed from the shared base states $h_j^{(s)}$; concatenation includes all heads. Only the query source changes between the two reads. Since each read has unit-sum weights within each head, their weight difference sums to zero and redistributes the same visual values. Appendices A.1–A.2 detail head-wise retrieval and cancellation.

3.4 Residual Write-back and Progressive Replay

At the receiving block b_s, CCR first computes the native post-attention residual U^{b_s}, adds $\lambda\Delta_i^{(s)}$ only at visual positions, and then applies the feed-forward transformation:

$$\begin{aligned} U^{b_s} &= H^{m_s} + \mathrm{MHA}_{b_s}\big(\mathcal{N}_{b_s}^A(H^{m_s}); M^{\mathrm{cau}}\big), \\ \widetilde{U}_i^{b_s} &= \begin{cases} U_i^{b_s} + \lambda\Delta_i^{(s)}, & i \in \mathcal{V}, \\ U_i^{b_s}, & i \notin \mathcal{V}, \end{cases} \qquad H^{b_s} = \widetilde{U}^{b_s} + F_{b_s}\Big(\mathcal{N}_{b_s}^F(\widetilde{U}^{b_s})\Big), \end{aligned} \tag{6}$$

where MHA_{b_s} is the native multi-head self-attention, including its frozen output projection; F_{b_s} is the frozen feed-forward network; and $\mathcal{N}_{b_s}^A$ and $\mathcal{N}_{b_s}^F$ are their respective input normalizations. The hyperparameter $\lambda \geq 0$ controls the correction strength and is shared across all S stages. Setting $\lambda = 0$ recovers the unmodified backbone under the same prompt. The temporary contextual states are discarded after the write-back.

For $s < S$, native blocks $b_s + 1, \ldots, m_{s+1} - 1$ process the updated sequence (Fig. 2(a)). Instruction states can read the corrected visual evidence from block $b_s + 1$ onward: the instruction text stays fixed, but its hidden states evolve. Stage $s + 1$ starts from $H^{m_{s+1}-1}$, recomputes the paired states, and rebuilds the shared visual memory from the new base states $H_j^{m_{s+1}}$, $j \in \mathcal{V}$. Its correction thus measures an additional contextual change against a base read shaped by earlier stages.

The stage count $S \in \mathbb{N}_+$ and ordered schedule $\mathbf{m} = (m_1, \ldots, m_S)$ of integer block indices are hyperparameters. With blocks indexed from 1 to L, admissible schedules satisfy $m_1 \geq 2$, $m_{s+1} \geq b_s + 2$ $(s = 1, \ldots, S-1)$, $b_S < L$.

These bounds leave at least one native block before the first stage, between successive stages, and after the final stage, allowing the decoder to form initial context, propagate corrections, and integrate the updated evidence into classification.

3.5 WHOLE-IMAGE EMOTION PREDICTION

After the final write-back, blocks $b_S + 1, \ldots, L$ complete native causal processing, allowing the classification suffix to integrate the updated visual evidence with the category descriptions. At the final prompt position t, including the backbone's assistant-prefix formatting, the unchanged output normalization and language-model head give vocabulary logits $z = W_{\mathrm{lm}}\mathcal{N}_{\mathrm{out}}(H_t^L) + b_{\mathrm{lm}}$. We select $\hat{k} = \arg\max_{k \in \{1,\ldots,K\}} z_{a_k}$ and return $\hat{y} = y_{\hat{k}}$, using the single-token identifiers in Sec. 3.1.

CCR runs in evaluation mode, encodes each image once, and preserves the original sequence length and visual-token positions. Each stage adds an auxiliary evaluation of the visual rows at m_s and two shared-memory reads at b_s; the remaining blocks follow the native computation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Architectures. We evaluate CCR on LLaVA-7B (Liu et al., 2023; 2024), VILA-8B (Lin et al., 2024), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025), with all model parameters frozen and no additional training or fine-tuning.

Datasets. We use four whole-image emotion recognition datasets: Emotion6 (Peng et al., 2015), EmoSet (Yang et al., 2023), WEBEmo (Panda et al., 2018), and Abstract (Machajdik & Hanbury, 2010). Evaluating WEBEmo at 7-category and 25-category granularities yields five settings. We report classification accuracy (ACC, %) and its unweighted mean across settings (Average). Gains are in percentage points; average gains are computed from per-setting ACC differences before rounding.

Implementation Details. Experiments use NVIDIA A100 80GB hardware and the inference protocol in Sec. 3, with a shared contextual instruction and dataset-specific category prompts, without image-specific human guidance. Unless otherwise specified, each backbone uses a fixed $(S, \mathbf{m}, \lambda^*)$ across images and all five settings. The CCR results in Tables 1 and 2 use $S = 3$, with $\lambda^* = 0.50, 1.00, 0.50$ for LLaVA-7B, VILA-8B, and Qwen2.5-VL-7B-Instruct, respectively. Section 4.4 examines stage count and correction strength.

4.2 MAIN RESULTS

Table 1 compares CCR with Zero-shot, Zero-shot-CoT (Kojima et al., 2022), and SEPM (Fang et al., 2025). CCR achieves the highest average ACC on all three backbones: 41.99% on LLaVA-7B, 41.77% on VILA-8B, and 47.02% on Qwen2.5-VL-7B-Instruct. Its average gains over SEPM, the strongest competitor by average ACC on each backbone, are 1.95, 3.40, and 3.06 points, respectively. CCR ranks first in 14 of the 15 backbone–setting combinations; the exception is LLaVA-7B on WEBEmo-7, where it obtains 41.38%, 1.01 points below SEPM. CCR improves over Zero-shot in all 15 combinations, including both WEBEmo granularities and every setting on Qwen2.5-VL-7B-Instruct, the backbone with the strongest initial average ACC. These results support its applicability across the evaluated backbones and category systems. The evaluated Zero-shot-CoT baseline yields lower average ACC than Zero-shot on every backbone despite occasional gains in individual settings, and does not reproduce CCR's consistent improvements. The following ablations examine the contributions of CCR's individual components.

4.3 ABLATION STUDY

Table 2 examines visual interventions, contextual feedback sources, and correction formulations, with Zero-shot and full CCR as shared references.

Table 1: Whole-image emotion recognition (ACC, %). Bold marks the best result per row and backbone. WEBEmo-7/25 are WEBEmo granularities; ZS/CoT denote Zero-shot/Zero-shot-CoT.

	LLaVA-7B				VILA-8B				Qwen2.5-VL-7B-Instruct			
Dataset	ZS	CoT	SEPM	CCR	ZS	CoT	SEPM	CCR	ZS	CoT	SEPM	CCR
Emotion6	48.32	48.65	54.21	**56.90**	47.47	44.28	51.35	**56.06**	58.42	55.56	59.76	**62.46**
EmoSet	52.77	51.67	56.04	**58.57**	44.45	43.20	53.11	**57.34**	55.98	55.36	58.73	**61.53**
WEBEmo-7	25.56	22.93	**42.39**	41.38	41.90	41.08	44.30	**46.75**	46.53	46.95	50.17	**52.73**
WEBEmo-25	15.71	15.52	18.26	**21.67**	15.35	13.57	17.36	**19.39**	22.54	22.97	24.00	**26.93**
Abstract	27.86	26.43	29.29	**31.43**	20.00	21.07	25.71	**29.29**	25.00	25.36	27.14	**31.43**
Average	34.04	33.04	40.04	**41.99**	33.83	32.64	38.37	**41.77**	41.69	41.24	43.96	**47.02**

Table 2: CCR ablations (ACC, %). A: alternative visual interventions; B: contextual feedback sources; C: correction formulations. Bold: best per column and backbone; Ref.: reference.

Group	Variant	Emotion6	EmoSet	WEBEmo-7	WEBEmo-25	Abstract	Average
LLaVA-7B							
Ref.	Zero-shot	48.32	52.77	25.56	15.71	27.86	34.04
A	Selection-only	51.68	54.73	**42.84**	17.41	28.93	39.12
A	Direct-context	51.18	54.54	28.60	17.58	28.21	36.02
B	Visual-only context	52.53	55.62	34.74	18.37	29.64	38.18
B	Image-blind instruction	52.19	56.98	36.88	19.25	30.00	39.06
C	State-diff	54.04	56.52	36.68	18.64	29.29	39.03
C	Full-read	52.53	56.96	35.27	19.78	29.29	38.77
Full	CCR	**56.90**	**58.57**	41.38	**21.67**	**31.43**	**41.99**
VILA-8B							
Ref.	Zero-shot	47.47	44.45	41.90	15.35	20.00	33.83
A	Selection-only	50.00	50.08	44.11	16.80	23.21	36.84
A	Direct-context	50.51	46.66	44.05	15.68	20.36	35.45
B	Visual-only context	52.19	49.49	44.63	17.04	24.29	37.53
B	Image-blind instruction	53.20	52.16	44.83	16.65	25.71	38.51
C	State-diff	53.20	52.29	44.97	17.91	26.07	38.89
C	Full-read	53.70	53.06	45.52	17.90	25.71	39.18
Full	CCR	**56.06**	**57.34**	**46.75**	**19.39**	**29.29**	**41.77**
Qwen2.5-VL-7B-Instruct							
Ref.	Zero-shot	58.42	55.98	46.53	22.54	25.00	41.69
A	Selection-only	59.43	58.21	50.50	23.30	25.71	43.43
A	Direct-context	59.43	58.53	48.67	23.73	25.00	43.07
B	Visual-only context	61.11	59.22	49.82	25.29	27.14	44.52
B	Image-blind instruction	61.28	58.73	49.70	24.11	27.50	44.26
C	State-diff	60.61	59.88	51.53	25.86	29.64	45.50
C	Full-read	60.10	58.43	50.13	25.25	28.93	44.57
Full	CCR	**62.46**	**61.53**	**52.73**	**26.93**	**31.43**	**47.02**

Alternative Visual Interventions. **Selection-only** selects visual tokens without CCR's representation correction, whereas **Direct-context** uses contextualized visual states directly instead of a differential retrieval update. Both improve average ACC over Zero-shot on all three backbones. CCR exceeds Direct-context in all 15 combinations, with average gains of 3.94–6.31 points across backbones. Its average gains over Selection-only are 2.87–4.93 points; the sole setting-level exception is LLaVA-7B on WEBEmo-7, where Selection-only obtains 42.84% versus CCR's 41.38%. Token selection remains competitive in this setting, while CCR achieves higher average ACC on every backbone. The Differential Retrieval paragraph below directly examines the correction formulation.

Source of Contextual Feedback. **Visual-only context** restricts feedback to visual information, whereas **Image-blind instruction** uses instruction information not conditioned on the input image. The latter concerns the auxiliary feedback source; the main branch still receives visual input. CCR outperforms both variants in all 15 combinations, with average gains of 2.50–4.24 points over Visual-only context and 2.75–3.26 points over Image-blind instruction. Neither restricted feedback source is uniformly preferable across backbones. CCR's consistent advantage supports its complete feedback configuration, combining visual context with image-conditioned instruction information.

Differential Retrieval. **State-diff** uses $r_i^{(s)} - h_i^{(s)}$, retaining subtraction but omitting visual retrieval. **Full-read** uses $\mathcal{R}_{s,i}(r_i^{(s)})$, retaining retrieval but omitting the base-state reference. CCR combines both elements through shared-memory differential retrieval (Eq. (4)), using the base read to measure the contextual change in retrieved evidence. It outperforms both variants in every combination, with average gains of 1.51–2.96 points over State-diff and 2.45–3.22 points over Full-read. These results support the complete formulation over either a direct hidden-state difference or the contextualized read alone. Section 4.4 examines progressive replay across decoder depths.

4.4 Progressive Replay and Correction Strength

Figure 3 examines replay stages and correction strength across three backbones and five settings.

Progressive replay. We test $S = 1, 2, 3$ by activating successive prefixes of (m_1, m_2, m_3), located at approximately one-quarter, one-half, and three-quarters of decoder depth, with receiving blocks $b_s = m_s + 1$. Thus, $S = 1$ uses m_1, whereas $S = 0$ disables replay. The analysis uses the fixed strengths in Sec. 4.1; all configurations finish the remaining native blocks before classification. **Progressive CCR** recomputes corrections from the updated main branch, whereas **Open-loop replay** writes back depth-specific corrections computed along the unmodified native trajectory. The two use matched replay positions, stage counts, and strengths, and coincide at $S = 1$.

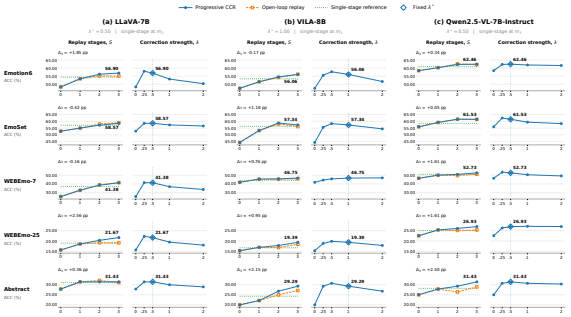


Figure 3: Progressive replay and correction strength. (a)–(c): three backbones; rows: five settings. Left: Progressive CCR (solid blue), Open-loop replay (dashed orange), and Single-stage reference (dotted green) at fixed λ^*. Horizontal references use one replay at m_2, m_2, m_3, respectively, independent of S. Right: λ sweeps at $S = 3$; diamonds mark fixed strengths from Tables 1 and 2, not curve maxima. Bold: three-stage CCR ACC (%) at λ^*; Δ_3: gain over Open-loop replay at $S = 3$ (percentage points).

Both replay variants improve mean ACC as stages are added. At $S = 3$, Progressive CCR exceeds Open-loop replay by 0.84, 0.97, and 1.22 points on LLaVA-7B, VILA-8B, and Qwen2.5-VL-7B-Instruct, respectively, and performs better in 12 of 15 combinations. This supports an additional benefit from recomputing corrections after earlier updates, beyond increasing write-back count alone. Under the tested schedule, Progressive CCR's mean ACC increases with each stage, with diminishing increments. At $S = 3$, it also exceeds the **Single-stage reference** in all 15 combinations, with mean gains of 2.24, 2.77, and 2.28 points, respectively. This reference uses one replay at m_2, m_2, m_3 for the three backbones, respectively, at their fixed strengths, and differs from prefix-based $S = 1$. Individual settings need not improve monotonically: VILA-8B on EmoSet decreases from 58.78% at $S = 2$ to 57.34% at $S = 3$.

Correction strength. With all three replay positions active, we vary $\lambda \in \{0, 0.25, 0.50, 1.00, 2.00\}$; $\lambda = 0$ recovers the unmodified backbone under the same prompt. Every tested positive strength improves ACC over $\lambda = 0$ in all 15 combinations, so the gains are not confined to a single strength. The highest observed mean occurs at $\lambda = 0.25, 0.50, 0.25$ for LLaVA-7B, VILA-8B, and Qwen2.5-VL-7B-Instruct, respectively. Increasing λ to 2.00 lowers mean ACC relative to these peaks on every backbone, although VILA-8B on WEBEmo-7 continues to improve through 2.00. The fixed CCR strengths used in Tables 1 and 2 yield mean ACC within 0.73 points of each backbone's highest observed mean. Figure 3 marks these fixed strengths rather than curve maxima. The sweep favors moderate strengths in aggregate without establishing a single optimum across backbones and settings. Together, these results distinguish the benefit of the tested multi-stage schedule from the additional contribution of state-dependent replay, while showing that increasing correction strength does not necessarily improve recognition.

4.5 Qualitative Analysis

Figure 4 shows final attention and stage-wise retrieval-weight differences for four examples of meaning ambiguity and limited representativeness; see Appendix B for more. Stage-s predictions use the first s replays followed by all remaining native decoder blocks.

Meaning ambiguity. In Fig. 4(a), CCR shifts relative attention from the lower performer's head toward the coordinated body configuration compared with the baseline, while the prediction changes from *fear* to *excitement*. The ropes and stage setting support interpreting the aerial pose as a performance. In Fig. 4(b), attention shifts from the toy's tense expression toward the barbell and grip, while the prediction changes from *anger* to *amusement*. The lifting action and toy context offer an interpretation of facial tension beyond anger alone.

Limited representativeness. In Fig. 4(c), attention shifts from the doll's face toward the dilapidated surroundings, while the prediction changes from *contentment* to *fear*; the face alone does not capture the scene's emotional character. In Fig. 4(d), relative attention shifts from the smiling adult toward the crying child, and the prediction changes from *contentment* to *sadness*. The adult's smile need not change its local meaning; its contribution to the whole-image judgment must instead be weighed against the child's contrasting expression.

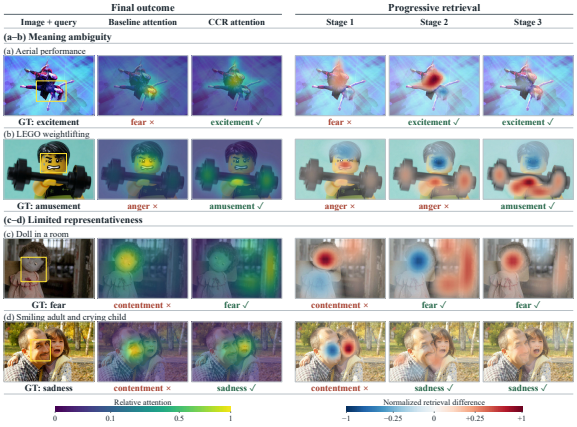


Figure 4: Qualitative analysis of contextual cue replay. Left: final relative visual attention and predictions of the baseline and CCR. Right: normalized stage-wise retrieval-weight differences (contextualized minus base); red/blue denote increased/decreased weights. Yellow boxes mark fixed query regions for visualization only; CCR operates on all visual positions. (a–b): meaning ambiguity; (c–d): limited representativeness. Stage-s predictions use the first s replays followed by all remaining native decoder blocks. GT: reference emotion label; ✓/×: correct/incorrect.

Progressive retrieval. Cases (a), (c), and (d) change to the reference label after two replays, whereas (b) requires three. Retrieval-weight and final-label changes need not coincide: in Fig. 4(b), Stage 2 reduces weight at the face and increases it around the lifting action, but the prediction changes to *amusement* only with three stages. Nor are retrieval-weight changes uniformly directed away from the queried cue: at the doll's face in Fig. 4(c), the difference is negative at Stage 2 and positive at Stage 3, while both configurations predict *fear*. Each difference uses a stage-specific base read and rebuilt visual memory, not a fixed cross-stage reference. These patterns are consistent with progressive contextual re-reading that adjusts how local cues contribute to whole-image judgments.

5 Conclusion

We presented Contextual Cue Replay (CCR), a training-free framework for whole-image emotion recognition that feeds task-conditioned scene information back into visual representations within frozen MLLMs, addressing the causal attention barrier to instruction-conditioned visual refinement. At selected decoder depths, CCR queries a shared visual memory with base and contextualized states and incrementally adds their retrieval difference to visual residuals. Native decoder blocks propagate corrections, letting later stages use updated visual evidence and instruction states. CCR retains all visual tokens without image-specific human guidance. Across three backbones and five settings, it achieves the highest average ACC among compared methods on each backbone, with average gains of 1.95–3.40 percentage points over SEPM. Ablation and replay analyses support image-conditioned contextual feedback, differential retrieval, and state-dependent replay, highlighting contextual refinement of visual evidence beyond token selection.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript drafting and polishing, figure refinement, and reproducing baseline implementations from prior work. These tools were not used to originate the core research ideas, design the experiments, or generate experimental data. All substantive research decisions and scientific conclusions were made by the authors. The authors reviewed the AI-assisted content and code and take full responsibility for the final manuscript and results.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923, 2025. URL https://arxiv.org/abs/2502.13923.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G. Hauptmann. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c7f43ada17acc234f568dc66da527418-Abstract-Conference.html.

Yiyang Fang, Jian Liang, Wenke Huang, He Li, Kehua Su, and Mang Ye. Catch Your Emotion: Sharpening Emotion Perception in Multimodal Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16029–16039. PMLR, 2025. URL https://proceedings.mlr.press/v267/fang25h.html.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

SangEun Lee, Yubeen Lee, Eunil Park, and Wonseok Chae. MASP: Multi-Aspect Guided Emotion Reasoning with Soft Prompt Tuning in Vision-Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(3):1882–1890, 2026. doi: 10.1609/aaai.v40i3.37168. URL https://ojs.aaai.org/index.php/AAAI/article/view/37168.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.16922.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On Pre-training for Visual Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26689–26699, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lin_VILA_On_Pre-training_for_Visual_Language_Models_CVPR_2024_paper.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2310.03744.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs. In *Computer Vision – ECCV 2024*, volume 15141 of *Lecture Notes in Computer Science*, pp. 125–140. Springer, 2025. doi: 10.1007/978-3-031-73010-8_8. URL https://link.springer.com/chapter/10.1007/978-3-031-73010-8_8.

Jana Machajdik and Allan Hanbury. Affective Image Classification Using Features Inspired by Psychology and Art Theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 83–92. Association for Computing Machinery, 2010. doi: 10.1145/1873951.1873965. URL https://doi.org/10.1145/1873951.1873965.

Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K. Roy-Chowdhury. Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias. In *Computer Vision – ECCV 2018*, volume 11206 of *Lecture Notes in Computer Science*, pp. 594–612. Springer, 2018. doi: 10.1007/978-3-030-01216-8_36. URL https://link.springer.com/chapter/10.1007/978-3-030-01216-8_36.

Xiaohuan Pei, Tao Huang, YanXiang Ma, and Chang Xu. Rethinking Causal Mask Attention for Vision-Language Inference. arXiv preprint arXiv:2505.18605, 2025. URL https://arxiv.org/abs/2505.18605.

Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C. Gallagher. A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 860–868, 2015. doi: 10.1109/CVPR.2015.7298687. URL https://doi.org/10.1109/CVPR.2015.7298687.

Saeedeh Sadeghi, Zijin Gu, Eve De Rosa, Amy Kuceyeski, and Adam K. Anderson. Direct Perception of Affective Valence from Vision. *Nature Communications*, 15:10735, 2024. doi: 10.1038/s41467-024-53668-6. URL https://www.nature.com/articles/s41467-024-53668-6.

Dong She, Xianrong Yao, Liqun Chen, Jinghe Yu, Yang Gao, and Zhanpeng Jin. AICA-Bench: Holistically Examining the Capabilities of VLMs in Affective Image Content Analysis. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13501–13528. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.661. URL https://aclanthology.org/2026.findings-acl.661/.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://arxiv.org/abs/2201.11903.

Daiqing Wu, Dongbao Yang, Sicheng Zhao, Can Ma, and Yu Zhou. Customizing Visual Emotion Evaluation for MLLMs: An Open-vocabulary, Multifaceted, and Scalable Approach. In *International Conference on Learning Representations*, pp. 96462–96482, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9bf12308ece130daa083fb21f7faf1b6-Abstract-Conference.html.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26596–26605, 2024. URL https://arxiv.org/abs/2404.16670.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On Layer Normalization in the Transformer Architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533. PMLR, 2020. URL https://proceedings.mlr.press/v119/xiong20b.html.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. EmoSet: A Large-scale Visual Emotion Dataset with Rich Attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20383–20394, 2023. URL https://arxiv.org/abs/2307.07961.

Qu Yang, Mang Ye, and Bo Du. EmoLLM: Multimodal Emotional Understanding Meets Large Language Models. arXiv preprint arXiv:2406.16442, 2024. URL https://arxiv.org/abs/2406.16442.

Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Wenjia Niu, Sabrina Caldwell, Tom Gedeon, Yang Liu, and Zhenyue Qin. Visual Prompting in LLMs for Enhancing Emotion Recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4484–4499. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.257. URL https://aclanthology.org/2024.emnlp-main.257/.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74840–74857. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25s.html.

A Detailed Computation of CCR

This appendix expands the shared-memory retrieval in Sec. 3.3 and derives the cancellation properties of its differential correction.

A.1 Head-wise Computation of Shared-Memory Retrieval

At stage s, the shared visual memory consists of the current main-branch base states $h_j^{(s)} = H_j^{m_s}$, $j \in \mathcal{V}$. Both reads reuse the receiving block b_s's frozen pre-attention normalization $\mathcal{N}_{b_s}^A$ and query, key, and value projections $Q_{b_s}^a$, $K_{b_s}^a$, and $V_{b_s}^a$ for head a, including any backbone biases. For query-source state x at visual position i, these operations give

$$\begin{aligned} q_{s,i}^a(x) &= \rho_i\big(Q_{b_s}^a\big(\mathcal{N}_{b_s}^A(x)\big)\big)\,, \\ k_{s,j}^a &= \rho_j\Big(K_{b_s}^a\Big(\mathcal{N}_{b_s}^A(h_j^{(s)})\Big)\Big)\,, \\ v_{s,j}^a &= V_{b_s}^a\Big(\mathcal{N}_{b_s}^A(h_j^{(s)})\Big)\,, \qquad j \in \mathcal{V}. \end{aligned} \tag{A.1}$$

Here ρ_i is the backbone's rotary transformation at the original position i. The backbone's head assignments and attention scaling are retained. The query source is $h_i^{(s)}$ for the base read and $r_i^{(s)}$ for the contextualized read; both reads use exactly the same keys and values.

For head dimension d_h, the attention weights are normalized over all visual positions:

$$\alpha_{s,ij}^a(x) = \frac{\exp\big(q_{s,i}^a(x)^\top k_{s,j}^a/\sqrt{d_h}\big)}{\sum_{u\in\mathcal{V}}\exp\big(q_{s,i}^a(x)^\top k_{s,u}^a/\sqrt{d_h}\big)}. \tag{A.2}$$

The projected retrieval output is

$$\mathcal{R}_{s,i}(x) = O_{b_s}\left(\mathrm{Concat}_a\left[\sum_{j\in\mathcal{V}}\alpha_{s,ij}^a(x)v_{s,j}^a\right]\right), \tag{A.3}$$

where concatenation includes all heads and O_{b_s} is the frozen attention-output projection, including its bias when present. Both reads attend over the complete visual memory, including visual positions later than i; neither is the native causally masked attention output. The memory is fixed within the paired reads and rebuilt from the new base states at each later stage.

A.2 Cancellation Properties of Differential Retrieval

For each head, define the difference between the contextualized and base attention weights. Since each read is normalized over the same visual positions,

$$\begin{aligned} \delta\alpha_{s,ij}^a &= \alpha_{s,ij}^a(r_i^{(s)}) - \alpha_{s,ij}^a(h_i^{(s)}), \\ \sum_{j\in\mathcal{V}}\delta\alpha_{s,ij}^a &= 1 - 1 = 0. \end{aligned} \tag{A.4}$$

Writing the output projection as $O_{b_s}(u) = W_{b_s}^O u + b_{b_s}^O$, the shared output-projection bias cancels when the two projected reads are subtracted. The correction is therefore

$$\begin{aligned} \Delta_i^{(s)} &= \mathcal{R}_{s,i}(r_i^{(s)}) - \mathcal{R}_{s,i}(h_i^{(s)}) \\ &= W_{b_s}^O\,\mathrm{Concat}_a\left[\sum_{j\in\mathcal{V}}\delta\alpha_{s,ij}^a v_{s,j}^a\right]. \end{aligned} \tag{A.5}$$

Any additive component c^a shared by all projected values in head a also contributes zero:

$$\sum_{j\in\mathcal{V}} \delta\alpha^a_{s,ij} c^a = \left(\sum_{j\in\mathcal{V}} \delta\alpha^a_{s,ij}\right) c^a = 0. \tag{A.6}$$

The correction thus measures a redistribution over the same visual values and vanishes whenever the two projected reads agree. These identities concern the fixed paired reads and their projected values. They do not establish invariance to changes in input hidden states that would also change normalization, keys, or queries.

B Additional Qualitative Examples

Figures B.1 and B.2 extend Sec. 4.5 with fifteen selected examples where CCR's final prediction matches the reference label and the baseline does not. They illustrate changes in visual emphasis and prediction across replay stages; aggregate performance is reported in Tables 1 and 2.

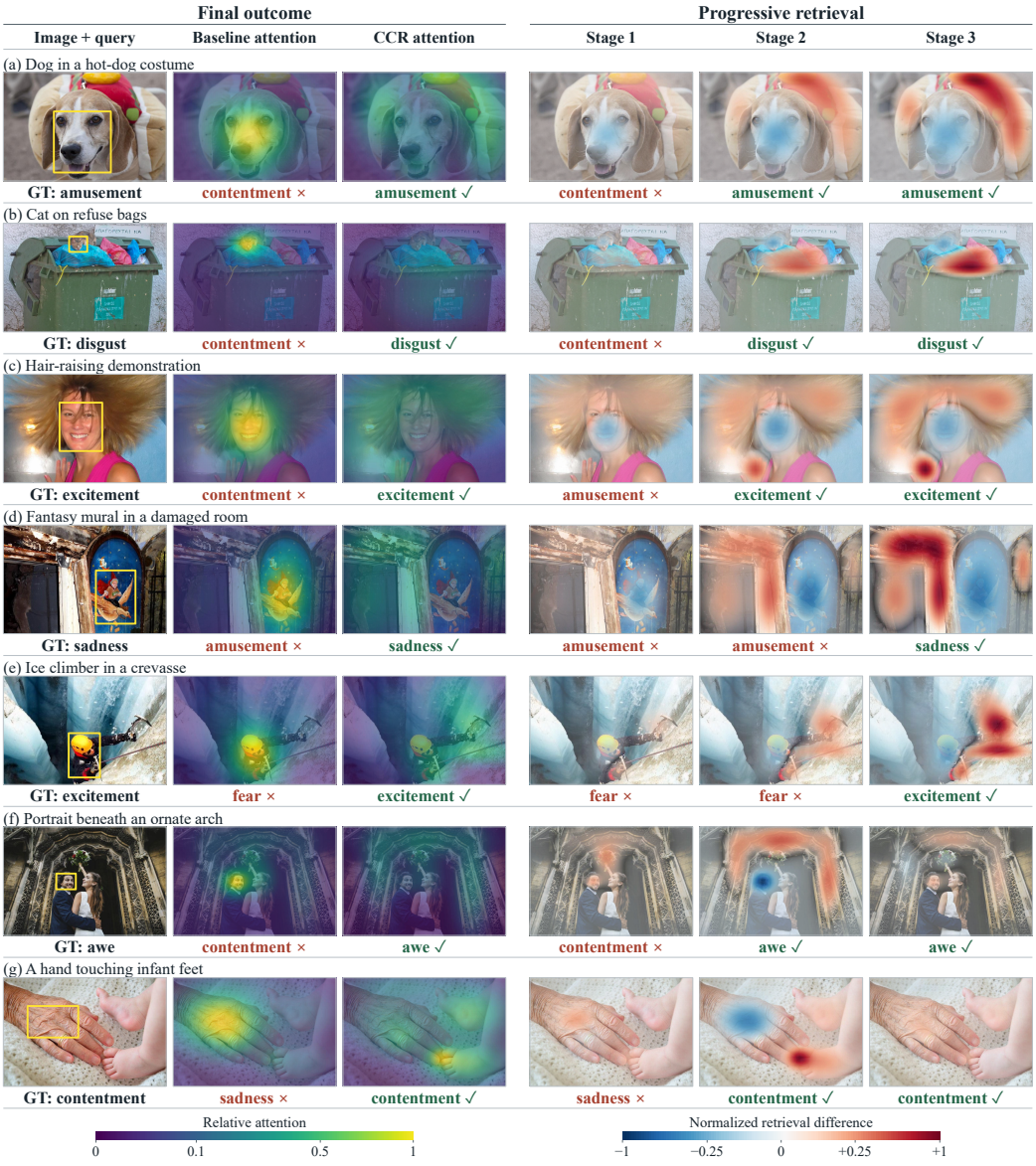


Figure B.1: Additional qualitative examples (a)–(g). Left: image, baseline and CCR final relative attention. Right: normalized retrieval-weight differences (contextualized minus base); red/blue indicate increases/decreases. Yellow boxes mark queries for visualization only; CCR uses all visual positions. Stage-s predictions use the first s replays followed by all remaining native decoder blocks. GT: reference label; ✓/×: correct/incorrect.

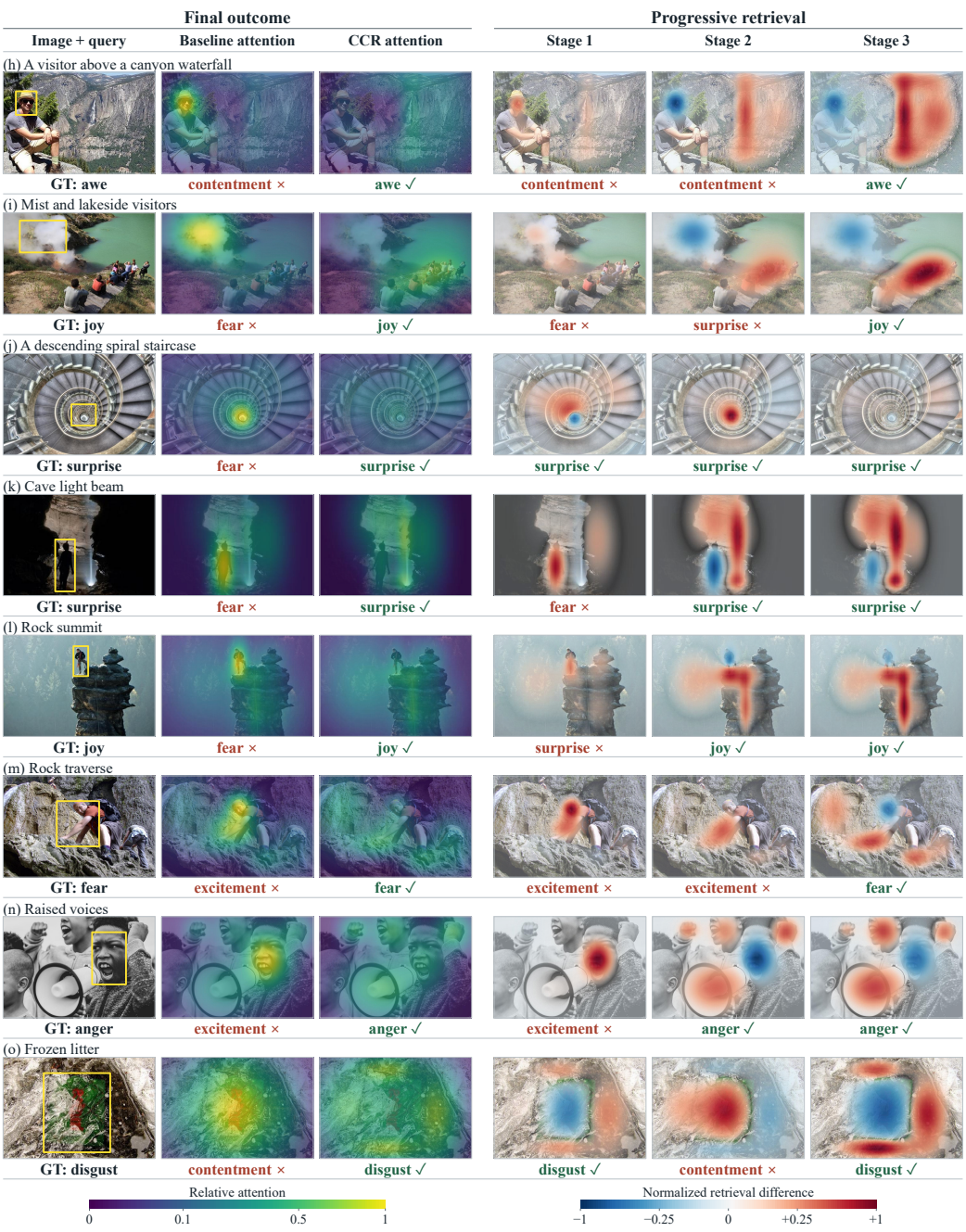


Figure B.2: Additional qualitative examples (h)–(o), using the same columns, color scales, query-box convention, and Stage-s prediction protocol as Fig. B.1.

Cases (a) and (d) illustrate how scene context can inform cue interpretation and contribution: the costume supports amusement in (a), while the damaged surroundings contrast with the mural in (d). Prediction trajectories need not improve monotonically: (j) reaches the reference label after one replay, whereas (o) changes from *disgust* to *contentment* and back to *disgust*. These selected trajectories complement the aggregate replay analysis in Sec. 4.4.