EMOPRUNE: LEARNING WHERE TO LOOK FOR MULTIMODAL EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) encode each image into hundreds of visual tokens, which make up most of the input sequence and thus most of the inference cost in multimodal emotion recognition (MER). Token pruning can reduce this cost, but for MER it is difficult to decide which tokens to keep: unlike visual question answering or grounding, an emotion prompt does not indicate where the evidence lies, and emotional cues are often small and not visually prominent. Generic pruning signals, which mainly reflect visual prominence, therefore tend to discard them. We propose EMOPRUNE, an emotion-aware visual token pruning method that learns which tokens to keep. A lightweight Evidence Head, trained offline from the frozen MLLM's answer loss with only image-level labels, scores visual tokens, and its scores are combined with `[CLS]` attention to retain both local cues and broader visual context. Unselected tokens are removed before the language model, without predefined regions or additional inference stages. Experiments on six benchmarks and two MLLM backbones show a favorable accuracy–compute trade-off. On LLaVA-1.5-7B, retaining 10% of visual tokens reduces computation from 8.8 to 1.78 TFLOPs per image while keeping accuracy close to full-token inference ($47.59 \pm 0.61\%$ vs. 47.81%).

1 INTRODUCTION

Multimodal large language models (MLLMs) show strong understanding and reasoning capabilities (Liu et al., 2023; Chen et al., 2024b; Wang et al., 2024). Beyond traditional vision–language tasks such as visual question answering (VQA) (Li et al., 2023; Dai et al., 2023; Fu et al., 2026), they are increasingly used for multimodal emotion understanding (Cheng et al., 2024; Lian et al., 2025; Hu et al., 2025). Multimodal emotion recognition (MER), a representative task in this domain, infers the emotion evoked by an image and supports applications such as human–computer interaction and content recommendation (Machajdik & Hanbury, 2010; Peng et al., 2015; Yang et al., 2023). Emotional evidence in an image is typically sparse: the decision often rests on a few small regions that may lie far apart, so much of the visual-token computation is spent on tokens that contribute little. This raises a question for efficient token pruning: which visual tokens should be retained?

Efficient token pruning for MER faces two selection challenges. ❶ **Emotional evidence is sparse and scattered.** Captioning focuses on global content, for which visually prominent regions usually suffice. In MER, the key information often lies in a few small, possibly distant regions, such as a facial expression, a hand gesture, or a small object, so the most prominent region is not necessarily the most informative one. ❷ **The prompt does not tell the model where to look.** A grounding prompt names the target (e.g., "What is the person holding?"), whereas an emotion prompt simply asks "What emotion does this image evoke?" and gives no spatial guidance. Figure 1 shows a motivating example at a 25% keep ratio: with all tokens or with tokens chosen by existing criteria, the compared methods make the same incorrect prediction. Efficient MER therefore depends not only on how many tokens are retained, but also on whether they preserve the task-relevant evidence.

Existing studies approach this problem from two directions. **Generic token pruning** keeps the top-ranked visual tokens under signals already inside the model, such as `[CLS]` attention, text–visual attention, or token diversity (Chen et al., 2024a; Zhang et al., 2024b; 2025; Yang et al., 2025; Fang et al., 2026). These signals do not measure emotional relevance, so they may keep prominent objects while discarding subtle emotion cues with the background. **Emotion-oriented evidence methods**

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.

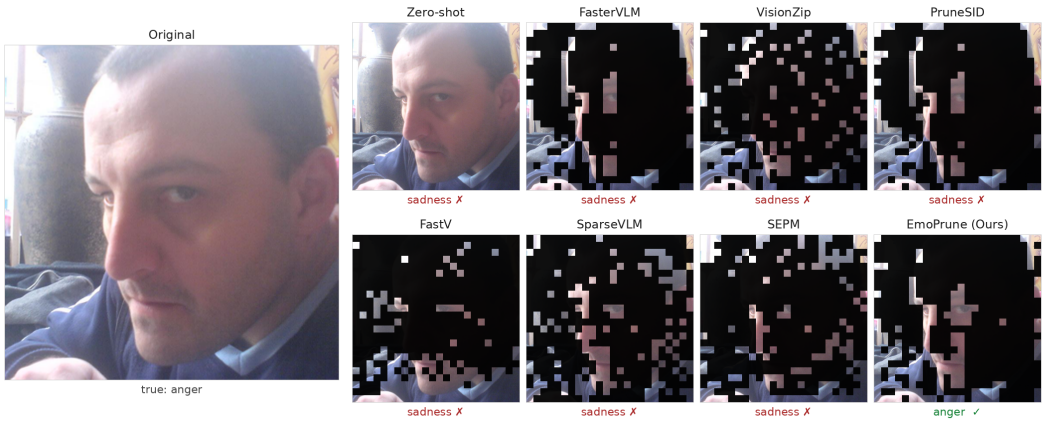


Figure 1: Comparison of visual token selection for MER at a 25% keep ratio. The original image is labeled "anger." Full-token inference and the compared selection methods predict "sadness," while EMOPRUNE retains the subtle facial expression and predicts "anger."

guide the MLLM toward candidate evidence by marking or segmenting regions, or by locating evidence in an additional forward pass (Zhang et al., 2024a; Fang et al., 2025; She et al., 2026). They rely on external region proposals or extra inference stages, and region-based designs assume that a few predefined regions capture the evidence, whereas sparse emotion cues may span several regions and become informative only in combination.

We propose EMOPRUNE, an emotion-aware visual token pruning method. The key idea is to let the frozen MLLM indicate, through its answer loss on labeled images, which visual tokens support the prediction, and to distill this into a lightweight scorer that runs alongside the vision encoder. Token-level selection retains sparse cues from distant regions without predefined regions, while `[CLS]` attention supplies subject and scene context. Because the backbone is frozen and the retained tokens enter the language model unchanged, accuracy differences at a fixed budget reflect only which tokens are kept. Our contributions are:

- We find that pruning signals designed for general vision–language tasks transfer poorly to emotion recognition: since an emotion prompt gives no spatial hint, they rank tokens mainly by visual prominence and tend to discard sparse emotion cues. This gap widens as the token budget tightens and is more pronounced in macro-F1 than in accuracy.
- We propose EMOPRUNE, which moves the evidence localization that existing emotion-oriented methods perform at inference, through extra passes or region proposals, into offline training. A lightweight head learns an emotion-specific token ranking from the frozen MLLM's answer loss with only image-level labels, so tokens can be pruned before the language model in a single pass.
- We evaluate EMOPRUNE on six emotion datasets with two MLLM backbones, where it achieves a favorable accuracy–compute trade-off under aggressive token reduction.

2 RELATED WORK

2.1 VISUAL TOKEN PRUNING

Visual token pruning removes tokens judged less useful for the output. Decoder-side approaches use the language model to guide selection: FastV (Chen et al., 2024a) ranks visual tokens by text–visual attention after the early decoder layers, and SparseVLM (Zhang et al., 2025) progressively removes tokens with text-related attention. Such rankings benefit from a question that names the relevant content, which a general emotion query does not provide, and part of the computation is paid before any reduction. Vision-side approaches select tokens before the language model: FasterVLM (Zhang et al., 2024b) ranks patches by `[CLS]` attention, VisionZip (Yang et al., 2025) combines dominant token selection with contextual token aggregation, and PruneSID (Fang et al., 2026) also considers

diversity to reduce redundancy. Unlike these fixed criteria, (Lei et al., 2026) train a small scorer for the ranking, but it reads the attention from the instruction to the image, so the ranking is still driven by the text prompt and inherits the same limitation when the prompt names no evidence. EMOPRUNE keeps the efficiency of vision-side pruning and adds an emotion-specific score, trained with the task loss on the emotion labels, to the selection criterion.

2.2 VISUAL EVIDENCE FOR EMOTION RECOGNITION

Emotion-oriented approaches guide MLLMs toward sparse visual evidence in faces, body poses, objects, or scene context. SoV (Zhang et al., 2024a) marks candidate regions in the input to direct attention, SEPM (Fang et al., 2025) locates evidence in one full pass and answers with the retained tokens in a second, and GAT (She et al., 2026) segments the image and verifies candidate regions step by step. These methods rely on marked regions, segmentation, or model-generated proposals rather than scoring individual tokens, and depend on external detectors or segmentation models, or on additional inference passes. EMOPRUNE instead scores visual tokens directly and selects evidence before the language model, without external region proposals or extra passes.

3 METHOD

3.1 TASK SETUP

An MLLM consists of a vision encoder f_{v}, a projector g, and a language model f_{L}. Given an image x and an instruction q containing the candidate emotion labels, the vision encoder extracts N patch features $F = f_{\mathrm{v}}(x)$. The projector maps each patch feature F_i to its corresponding visual token V_i, giving $V = g(F)$ for $i = 1, \ldots, N$. The visual and text tokens are then concatenated and processed by f_{L} to predict an emotion label $\hat{y} = f_{\mathrm{L}}([V, q])$, where $\hat{y}$ is the prediction and $y \in \mathcal{Y}$ is the ground-truth label from the label set $\mathcal{Y}$. Our goal is to retain a compact subset of visual tokens that preserves the evidence needed for emotion recognition. Let K denote the inference token budget and $\mathcal{I}$ the indices of the K selected tokens. The pruned sequence is $V_{\mathcal{I}}$, and the corresponding prediction is $\hat{y} = f_{\mathrm{L}}([V_{\mathcal{I}}, q])$. Section 3.5 specifies K from the drop rate.

3.2 OVERVIEW

Figure 2 presents the overall design of EMOPRUNE. The frozen vision encoder produces features $F = f_{\mathrm{v}}(x)$. The Evidence Head h_ψ, parameterized by ψ, assigns an emotion-specific score to each visual token, and the vision encoder provides a complementary `[CLS]`-attention score. Only h_ψ is trained; on a 7B backbone it has 0.87M parameters, under 0.02% of the full model. During training, a soft mask modulates each visual token so that task supervision reaches the head; at inference, the two scores rank the tokens and the top-K tokens are passed to the language model. We describe the Evidence Head (Section 3.3), training objective (Section 3.4), and inference (Section 3.5) below.

3.3 EVIDENCE HEAD

The Evidence Head h_ψ takes as input the patch features F from the vision encoder f_{v} and outputs one scalar score per patch. We first add a learnable positional embedding P, which has the same shape as F, and then apply a shared three-layer MLP with GELU activations:

$$s = h_\psi(F) = \mathrm{MLP}_\psi(F + P) \in \mathbb{R}^N. \tag{1}$$

Here, s is the vector of token scores and MLP_ψ denotes the MLP parameterized by ψ.

3.4 TRAINING OBJECTIVE

Task supervision with a soft mask. We train the Evidence Head to identify visual tokens that are relevant to the target answer. We convert the score vector s into a soft mask m using the sigmoid function σ:

$$m = \sigma(s) \in (0, 1)^N, \tag{2}$$

where m_i is the soft importance of the i-th visual token. We apply the mask to each visual token V_i:

$$\tilde{v}_i = m_i V_i, \qquad i = 1, \ldots, N, \tag{3}$$

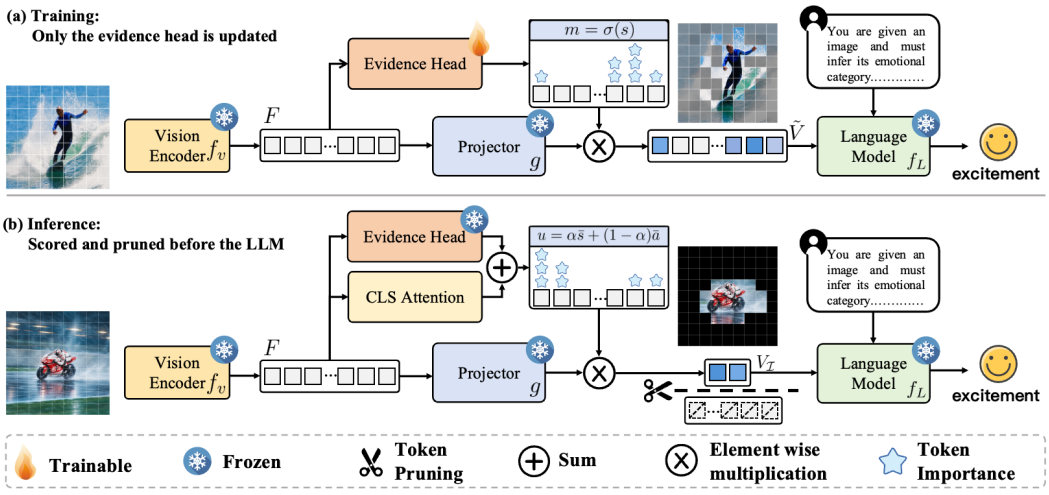


Figure 2: Training and inference pipeline of EMOPRUNE. The fire and snowflake icons mark trainable and frozen modules, respectively.

The masked tokens form $\tilde{V}$, and the frozen language model f_{L} receives $[\tilde{V}, q]$. Let $(y_1, \ldots, y_T)$ be the sequence of T tokens obtained by tokenizing the ground-truth label y. We define the task loss as the autoregressive cross-entropy over these answer tokens:

$$\mathcal{L}_{\text{task}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_{f_{\mathrm{L}}}\left(y_t \mid [\tilde{V}, q], y_{<t}\right), \tag{4}$$

where $y_{<t}$ denotes the preceding answer tokens and $p_{f_{\mathrm{L}}}$ is the next-token distribution of f_{L}; prompt and visual-token positions are excluded. The loss updates ψ through m without modifying f_{v}, g, or f_{L}. Soft masking keeps the sequence length and token positions unchanged, although scaling by m_i changes token norms.

From soft masking to hard selection. During training, all N tokens are kept and weighted by m_i, which makes selection differentiable. At inference, the scores are used only for ranking: the top-K tokens keep their original, unweighted embeddings, and the rest are discarded (Section 3.5). The soft mask thus serves as a differentiable surrogate for ranking rather than a test-time weighting, and the budget regularizer below makes the scores selective enough for top-K selection.

Budget regularizer and total objective. The task loss encourages retaining any visual token that may help predict the answer, but does not penalize irrelevant tokens. The mask may therefore assign large values to most patches, leaving the scores unable to distinguish which patches should be retained when only a small subset is used at inference. Let k^* denote the target number of retained tokens during training. We define the budget loss $\mathcal{L}_{\text{bud}}$ as

$$\mathcal{L}_{\text{bud}} = \left(\frac{1}{N}\sum_{i=1}^{N} m_i - \frac{k^*}{N}\right)^2. \tag{5}$$

It pulls the average mask value toward the target keep ratio without forcing every image to retain exactly k^* patches. k^* is used only in training; at inference, the number of retained tokens is set by the drop rate β. The total training objective $\mathcal{L}$ is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{bud}}. \tag{6}$$

3.5 INFERENCE

At inference, we combine the scores s from the Evidence Head with the `[CLS]` attention scores to rank the visual tokens. Both signals come from the same vision-encoder forward pass. With H

attention heads, the [CLS] score a_i of the i-th patch is the head-averaged attention weight:

$$a_i = \frac{1}{H}\sum_{j=1}^{H} A_j[\texttt{[CLS]}, i], \qquad i = 1, \ldots, N, \tag{7}$$

where A_j is the attention matrix of the j-th attention head, and i indexes the image patches after excluding the [CLS] token. Let $a = (a_1, \ldots, a_N)$ denote the resulting score vector. Since s and a have different scales, we standardize each vector over its N entries within an image:

$$\bar{s} = \frac{s - \mathrm{mean}(s)}{\mathrm{std}(s)}, \qquad \bar{a} = \frac{a - \mathrm{mean}(a)}{\mathrm{std}(a)}. \tag{8}$$

We then compute the combined score vector u:

$$u = \alpha\,\bar{s} + (1 - \alpha)\,\bar{a}, \tag{9}$$

where $\alpha \in [0, 1]$ controls the contribution of the Evidence Head. Setting $\alpha = 0$ uses only [CLS] attention, and $\alpha = 1$ uses only the head scores; we use $\alpha = 0.5$ (Section 4.4). Given a drop rate $\beta \in [0, 1)$, we retain the K visual tokens with the highest combined scores, where

$$K = \max\big(1, \mathrm{round}((1 - \beta)N)\big). \tag{10}$$

The top-K ranking determines $\mathcal{I}$; we sort these indices in ascending order so that the retained tokens remain in their original raster-scan order. We pass $V_{\mathcal{I}}$ to the frozen language model together with the instruction q:

$$\hat{y} = f_{\mathrm{L}}([V_{\mathcal{I}}, q]). \tag{11}$$

Tokens outside $\mathcal{I}$ never enter the language model, and the prediction is produced in a single pass.

4 Experiments

4.1 Setup

We evaluate on six image emotion datasets with 6 to 25 classes: FI, Emotion6, ArtPhoto, EmoSet, UnbiasedEmo, and WebEmo$_{25}$. For each of LLaVA-1.5-7B and InternVL3-8B, a separate Evidence Head is trained on 2,000 class-balanced EmoSet training images with the backbone frozen, selected on a validation pool from the non-test splits of four test sources, and applied to all six test sets. We report accuracy and per-image TFLOPs; training details and baselines are in Appendix A.

4.2 Comparison with Selection Strategies

Table 1 compares EmoPrune with full-token inference (No pruning) and token-selection baselines on two frozen backbones. Pruning methods are evaluated at matched token keep ratios.

Cost. EmoPrune prunes before the language model and adds only a small head, so its cost matches other vision-side methods (1.78/1.39 TFLOPs at 10%/5% on LLaVA-1.5-7B vs. 8.8 for full tokens); decoder-side methods cost more, and two-pass SEPM exceeds full-token inference.

Accuracy. On LLaVA-1.5-7B, EmoPrune achieves the best average accuracy among pruning methods at both budgets. At 10%, it matches full-token inference (47.59 ± 0.61 vs. 47.81) and leads the strongest baseline, PruneSID, by 1.0 point; at 5%, its margin over the second-best SparseVLM grows to 2.2 points. On InternVL3-8B, it is the best at 10% (49.29 vs. 48.55 for PruneSID); at 5%, it leads all single-pass methods by 1.5 points and stays within 0.2 points of two-pass SEPM at under a third of its cost. On both backbones, the margin over single-pass baselines widens as the budget tightens, consistent with the scoring-signal ablation in Section 4.3.

Macro-F1. Since accuracy can be dominated by frequent emotions, we also report macro-F1 in Appendix B.1. EmoPrune achieves the best average macro-F1 in all four pruning settings, with a larger margin than under accuracy, whereas some baselines that are competitive in accuracy, such as SEPM, fall sharply.

Table 1: Comparison of accuracy (%) and per-image cost on two models. Accuracy is reported on six datasets; Avg. is their unweighted mean. TFLOPs/img is the per-image mean. No pruning uses all visual tokens; for pruning methods, β is the drop rate (β=0.9 keeps 10% and β=0.95 keeps 5%). In the Avg. column, bold is best in each block and underline is second best. EMOPRUNE reports mean±std over three independent retrainings (data resampling and initialization).

		Function	FI 8	Emotion6 6	ArtPhoto 8	EmoSet 8	UnbiasedEmo 6	WebEmo_{25} 25	Avg.	TFLOPs/img↓
LLaVA-1.5-7B		No pruning	58.90	34.51	39.70	54.13	83.20	16.40	47.81	8.8
	β=0.9	SEPM	43.29	34.51	38.09	42.20	43.80	22.40	37.38	12.0
		FastV	50.45	20.54	33.50	42.83	56.10	15.10	36.42	2.2
		SparseVLM	60.11	32.83	39.33	52.74	77.40	13.60	46.00	2.7
		FasterVLM	55.97	38.05	38.71	49.08	81.40	14.00	46.20	1.78
		VisionZip	54.38	33.84	37.34	47.14	77.20	11.60	43.58	1.78
		PruneSID	59.98	33.33	41.19	51.11	81.30	12.60	<u>46.59</u>	1.78
		EMOPRUNE	55.58±0.52	40.85±0.42	40.07±1.33	51.92±1.52	84.13±0.09	12.97±0.38	**47.59**±0.61	1.78
	β=0.95	SEPM	33.12	28.11	30.15	34.67	42.50	19.20	31.29	11.6
		FastV	35.88	18.01	29.03	30.58	37.80	11.40	27.12	1.85
		SparseVLM	56.58	28.79	37.34	49.93	73.30	11.40	<u>42.89</u>	2.2
		FasterVLM	51.10	33.33	35.73	45.22	76.30	10.90	42.10	1.39
		VisionZip	48.81	27.10	32.63	41.49	68.40	8.30	37.79	1.39
		PruneSID	56.53	26.60	38.83	45.96	76.20	9.40	42.25	1.39
		EMOPRUNE	53.41±1.13	36.59±0.29	38.63±0.82	48.63±2.20	82.90±0.37	10.47±0.83	**45.10**±0.49	1.39
InternVL3-8B		No pruning	62.18	53.87	49.63	57.99	83.10	17.90	54.11	5.7
	β=0.9	SEPM	53.99	48.32	43.42	55.55	70.40	17.90	48.26	6.8
		FasterVLM	52.65	47.31	42.68	53.37	77.00	15.20	48.03	2.2
		VisionZip	55.37	46.46	42.43	54.78	75.40	16.40	48.47	2.2
		PruneSID	54.72	48.15	42.80	54.63	74.70	16.30	<u>48.55</u>	2.2
		EMOPRUNE	53.28±0.84	49.38±0.16	44.83±0.97	54.73±0.77	74.90±1.95	18.63±1.43	**49.29**±0.88	2.2
	β=0.95	SEPM	51.10	45.29	41.69	52.37	63.50	14.20	**44.69**	6.7
		FasterVLM	42.73	35.35	36.60	44.81	66.60	12.30	39.73	2.0
		VisionZip	49.42	41.58	36.97	50.11	68.20	11.90	43.03	2.0
		PruneSID	47.95	41.25	36.48	48.49	68.80	12.70	42.61	2.0
		EMOPRUNE	46.63±1.01	45.17±1.14	40.12±1.42	48.74±1.00	68.47±3.20	17.90±0.62	<u>44.50</u>±1.37	2.0

Table 2: Ablation on scoring signals across token budgets with frozen LLaVA-1.5-7B. Entries are mean accuracy (%) over the six test sets. ↑/↓ is relative to the No pruning mean (47.81); bold is best per budget. Head ⊕ [CLS] denotes the combined score in equation 9. Results are from a single run; Table 1 reports the three-seed mean of the full method.

Scoring signal	**Keep 50%**	**Keep 25%**	**Keep 15%**	**Keep 10%**	**Keep 5%**
Random top-k	47.61↓0.20	46.08↓1.73	44.28↓3.53	42.07↓5.74	37.55↓10.26
EMOPRUNE ([CLS])	47.61↓0.20	47.29↓0.52	46.40↓1.41	45.21↓2.60	41.81↓6.00
EMOPRUNE (Head)	47.60↓0.21	47.77↓0.04	47.90↑0.09	47.46↓0.35	45.70↓2.11
EMOPRUNE (Head ⊕ [CLS])	**47.67**↓0.14	**47.90**↑0.09	**48.50**↑0.69	**48.40**↑0.59	**45.80**↓2.01

4.3 ABLATION ON SCORING SIGNALS

We analyze the scoring signals to compare the Evidence Head with [CLS] attention and random selection under different token budgets, as shown in Table 2. When many tokens are retained, the choice of signal has a limited effect, but random selection degrades rapidly as the budget becomes smaller. The Evidence Head is generally stronger than [CLS] under aggressive pruning: at the 10% budget, it reaches 47.46 compared with 45.21 for [CLS], and at the 5% budget it reaches 45.70, exceeding [CLS] by 3.89 points. This suggests that the learned head preserves emotion-relevant evidence better than generic [CLS] saliency when few tokens remain.

We further study whether combining the two signals helps. The combined score achieves the best average accuracy at every budget, including 25% (47.90), 15% (48.50), 10% (48.40), and 5% (45.80). Its gain over the head alone is largest at 15% and 10% (0.60 and 0.94 points), consistent with [CLS] adding subject and scene context to the sparse cues selected by the head.

Table 3: Loss ablation: mean accuracy (%) on Emotion6 and ArtPhoto over keep $\{50, 25, 10, 5\}\%$. Rows with a head differ only in the loss and use the combined score (α=0.5).

Losses	Task	Bud.	E6	AP	Avg.
`[CLS]` only	—	—	36.57	37.69	37.13
budget only		✓	37.04	37.81	37.43
task only	✓		37.16	38.65	37.91
Task + budget	✓	✓	**37.84**	**40.20**	**39.02**

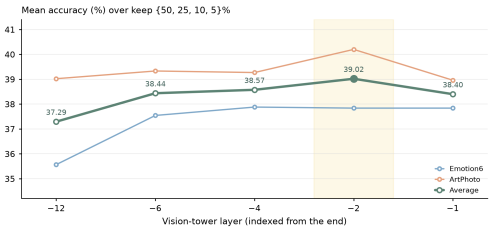


Figure 3: Layer ablation for the Evidence Head input on Emotion6 and ArtPhoto. Shaded: the layer LLaVA feeds to the LLM.

Table 4: Combination weight α: mean accuracy (%) on the validation pool over keep $\{50, 25, 10, 5\}\%$.

α	0	0.25	0.50	0.75	1
EmoSet	49.95	52.75	54.45	55.50	**56.80**
FI	54.40	55.20	56.95	56.90	**57.65**
Emotion6	**36.95**	36.80	36.50	35.35	32.20
WebEmo_{25}	8.70	9.00	**9.35**	9.20	8.95
Avg.	37.50	38.44	**39.31**	39.24	38.90

Table 5: Training budget k^*: mean accuracy (%) on the validation pool over keep $\{50, 25, 10, 5\}\%$.

k^*	$N/8$	$N/4$	$N/2$
EmoSet	52.60	**54.45**	51.50
FI	55.25	**56.95**	53.40
Emotion6	**37.85**	36.50	36.85
WebEmo_{25}	8.65	**9.35**	8.95
Avg.	38.59	**39.31**	37.67

4.4 Ablation Studies

All ablations use LLaVA-1.5-7B, and the configuration is transferred to InternVL3-8B without further tuning. Loss composition and feature layer are analyzed on the Emotion6 and ArtPhoto test sets, while k^* and α are selected on the validation pool. All ablations are single runs.

Loss composition. Table 3 isolates the training objective: every variant with a head is evaluated with the deployed combined score, so changes in the loss are not confounded with changes in the scorer, and `[CLS]` only serves as the reference without a head. With the budget term alone, the head receives no signal from the answer and stays close to this reference (+0.3 points). The full objective performs best; removing the task term or the budget term lowers the average by 1.6 and 1.1 points, respectively. We attribute the first drop to losing the discriminative signal from the answer, and the second to less selective scores for top-K ranking. Both terms are therefore needed.

Hyperparameter selection. The budget target k^* and the combination weight α are selected on the held-out validation pool (Tables 4 and 5). For k^*, $N/4$ gives the best average, 0.7 and 1.6 points above $N/8$ and $N/2$. For α, the sweep covers five values from `[CLS]`-only to head-only. Both single-signal settings fall below $\alpha = 0.5$ and 0.75 on the average, and the head alone drops on Emotion6, where `[CLS]` context helps. The equal weight $\alpha = 0.5$ gives the best average (39.31), with $\alpha = 0.75$ close behind (39.24). We use $k^* = N/4$ and $\alpha = 0.5$ throughout.

Feature layer. Scoring emotional content requires semantic features, and late ViT layers are more global and object-level (Raghu et al., 2021; Park & Kim, 2022). We therefore fix the head's input in advance to the penultimate layer (-2), which is also the layer LLaVA passes to the language model. Figure 3 reports a post-hoc sweep from -12 to -1: the middle layer is 1.7 points lower on the average, accuracy rises toward late layers, and -2 performs best, while the final layer is lower again. The gap between layers also widens as the budget tightens: at keep 50% all late layers stay within 0.4 points, whereas at keep 5% the penultimate layer leads by more than one point.

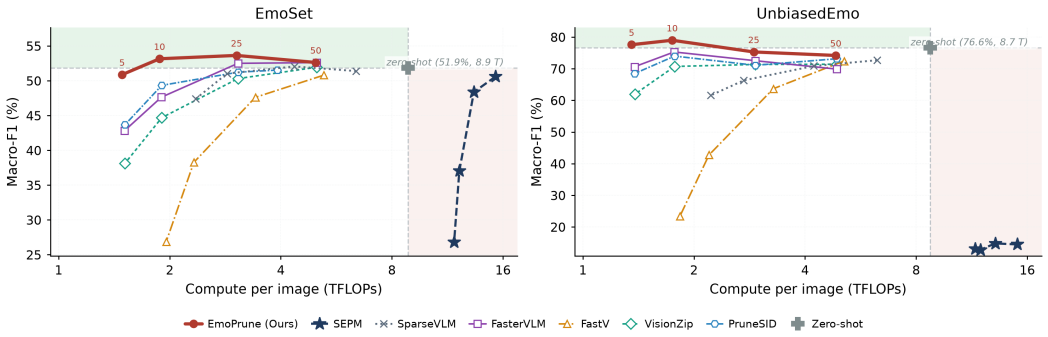


Figure 4: Macro-F1 versus per-image compute on EmoSet (left) and UnbiasedEmo (right) with LLaVA-1.5-7B, the cross marking the No pruning reference.

4.5 Efficiency and Pareto Analysis

Figure 4 plots macro-F1 against per-image compute on EmoSet and UnbiasedEmo with LLaVA-1.5-7B at keep ratios of 5%, 10%, 25%, and 50%. On both datasets, EMOPRUNE forms the upper envelope across the compute range and stays close to or above the full-token level. As compute decreases, the baselines lose macro-F1 quickly, whereas EMOPRUNE declines only slightly. SEPM costs more than full-token inference yet remains well below it. Curves for the other four datasets are in Appendix B.2.

4.6 [CLS] Attention versus Evidence Head

Figure 5 compares the regions retained by [CLS] attention, the Evidence Head, and their combination. Under the same budget, the selections differ, so the learned scores do not simply reproduce the [CLS] ranking. In the first example, [CLS] keeps the people and snowy ground, while the head selects more of the sky; the combination predicts "excitement" correctly although the head alone does not. In the second, the selections differ around the face: [CLS] predicts "sadness," whereas the head and the combination predict "anger" correctly. In the third, both signals predict "anger," but their combination predicts "sadness" correctly. These cases qualitatively support complementarity rather than accurate emotion localization; Appendix C gives more examples.



Figure 5: From left to right, the panels show the original image, the selections from [CLS] attention, the trained Evidence Head, and their combination.

4.7 Case Study

Figure 6 compares the evidence retained by different token-selection methods. Under the same budget, the methods preserve substantially different regions. EMOPRUNE tends to keep localized cues such as facial expressions, body interactions, and emotion-relevant objects, while other methods

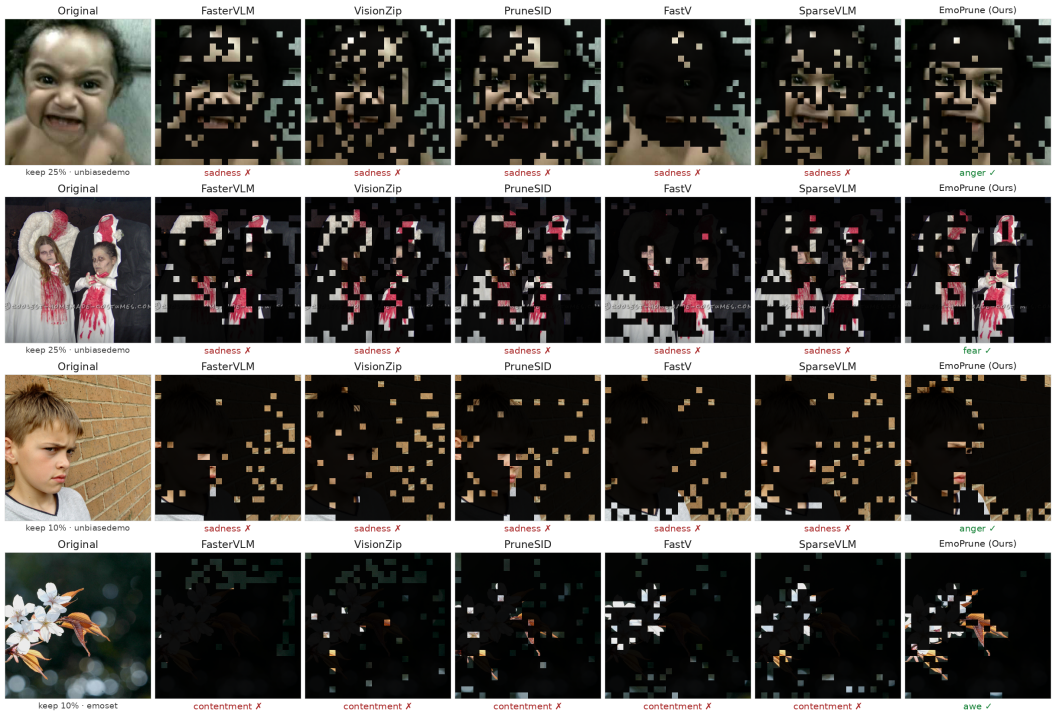


Figure 6: Comparison of visual token selection at 10% (top two rows) and 25% (bottom two rows) keep ratios. Each row shows the original image followed by the tokens retained by the compared methods and EMOPRUNE; the text below each panel gives its prediction. Black patches are discarded. In these examples, the compared methods often preserve fragments of the main subject but miss local emotion cues, whereas EMOPRUNE retains evidence that supports the correct prediction.

may retain broader regions and miss these details. The examples suggest that retaining the main subject alone may not preserve the local details needed to distinguish emotions.

5 LIMITATIONS

The combination weight is fixed across budgets; when the two signals favor different regions, a small budget may preserve neither completely (Appendix D), and adapting the weight to the budget is a natural extension. In addition, the head scores each patch independently and is therefore suited to sparse evidence, whereas emotion conveyed by the whole scene may call for token interaction, which we leave for future work.

6 CONCLUSION

We presented EMOPRUNE, a lightweight method for multimodal emotion recognition that learns which visual evidence is useful while keeping the MLLM backbone frozen. The method adds a small offline training cost and uses the learned emotion-specific ranking to prune visual tokens before they reach the language model, avoiding the extra inference stages required by many fine-tuning-free emotion methods. Across six benchmarks and two frozen MLLM backbones, EMOPRUNE maintains a favorable balance between accuracy and computation under aggressive token reduction. On LLaVA-1.5-7B, for example, it keeps 10% of the visual tokens, reduces the per-image cost by 4.9×, and achieves accuracy comparable to full-token inference in our evaluation. These results show that learning which visual evidence to retain can make frozen MLLMs more efficient for emotion recognition without modifying the backbone.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing polishing and language refinement, including improving clarity and readability of the manuscript. We also used generative AI tools for retrieval and discovery purposes, such as assisting the identification and exploration of related work. All AI-assisted outputs were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, and references.

REPRODUCIBILITY STATEMENT

The Evidence Head architecture, training objective, and inference procedure are described in Section 3. Appendix A lists the sources and label spaces of all six test sets, the construction of the training and validation pools, the public backbone checkpoints, the optimizer and hyperparameters, the checkpoint-selection rule, and the FLOPs accounting used for every method. All reported multi-seed results average three independent runs. Code for training and evaluation will be released upon acceptance.

REFERENCES

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

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. *Advances in Neural Information Processing Systems*, 37: 110805–110853, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

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 *Forty-second International Conference on Machine Learning*, 2025.

Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Prune redundancy, preserve essence: Vision token compression in VLMs via synergistic importance-diversity. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=i36E5Ezm0H.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

He Hu, Lianzhong You, Hongbo Xu, Qianning Wang, Fei Richard Yu, Fei Ma, Zebang Cheng, Zheng Lian, Yucheng Zhou, and Laizhong Cui. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*, 2025.

Lei Lei, Jie Gu, Xiaokang Ma, Chu Tang, Jingmin Chen, and Tong Xu. Task-related token compression in multimodal large language models from an explainability perspective. In *International Conference on Learning Representations*, volume 2026, pp. 112724–112746, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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, 2010.

Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

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*, pp. 860–868, 2015.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

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, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny 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*, pp. 20383–20394, 2023.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025.

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, 2024a.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, MinQi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv preprint arXiv:2412.01818*, 2024b.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning*, 2025.

A Experimental Details

A.1 Datasets and Data Splits

Test benchmarks. The six test sets are FI (2,319 images), Emotion6 (594), ArtPhoto (806), EmoSet (17,716), UnbiasedEmo (1,000), and WebEmo_{25} (1,000). Their label spaces contain 8, 6, 8, 8, 6, and 25 emotion categories, respectively. All datasets are publicly available: FI, Emotion6, and ArtPhoto are obtained from a public collection of affective image datasets, while the UnbiasedEmo and WebEmo_{25} test splits follow the VECBench release.

Training and validation splits. We construct the training and validation pools separately. The training pool contains 2,000 images sampled from the EmoSet training split, with 250 images per class and random sampling within each class. The validation pool contains 2,000 images, with 500 images randomly sampled per domain from the official non-test splits of the four tested sources: EmoSet, FI, Emotion6, and WebEmo$_{25}$. The training pool, the validation pool, and all test sets are pairwise image-disjoint.

A.2 Backbones and Compared Methods

Backbones. We use the publicly released checkpoints of LLaVA-1.5-7B and InternVL3-8B.

Backbone-specific Evidence Heads. We independently train an Evidence Head for each MLLM backbone. For both backbones, the head input and the `[CLS]` attention in equation 7 are taken from the penultimate vision-encoder layer. Both heads use the same three-layer MLP design, while their learnable positional embeddings P match the spatial layout of the corresponding vision encoder. For LLaVA-1.5-7B, the head processes a 24×24 grid of 576 patch features and contains 0.87M parameters. For InternVL3-8B, it processes a 32×32 grid of 1,024 patch features and contains 1.33M parameters, and its predicted patch scores are aggregated by 2×2 max pooling to align with the 256 visual tokens produced by the backbone's pixel-shuffle operation.

Compared methods. We compare full-token inference without pruning, the emotion-oriented method SEPM (Fang et al., 2025), and token-pruning baselines. SEPM first locates evidence with the MLLM and then answers from the retained tokens in a second pass. FastV (Chen et al., 2024a) and SparseVLM (Zhang et al., 2025) remove tokens inside the language model, while FasterVLM (Zhang et al., 2024b), VisionZip (Yang et al., 2025), and PruneSID (Fang et al., 2026) reduce visual tokens before they enter the language model. Random selection is included as a reference in the ablation study. Selection methods are evaluated at matched keep ratios whenever applicable.

A.3 Training and Evaluation Protocol

Training configuration and checkpoint selection. For each backbone, we train only its Evidence Head and keep the remaining parameters frozen. We optimize the head with AdamW using a learning rate of 10^{-4}, a batch size of 2, and four gradient-accumulation steps, giving an effective batch size of 8. The training objective is the full task-plus-budget loss with budget target $k^* = N/4$, and the combination weight is fixed to $\alpha = 0.5$; both are selected on the held-out validation pool. After each epoch we evaluate the validation pool, retain the checkpoint with the lowest validation cross-entropy, and stop at the first subsequent epoch without improvement. Each independent run uses a new class-balanced training sample and a new parameter initialization; the reported multi-seed results average three independent runs. Per-image FLOPs follow one rule for every method: the vision-encoder forward pass, LLM prefill and decoding, and all method-specific computation. All experiments are conducted on one RTX 5090 GPU.

B Macro-F1 Results

B.1 Comparison with Selection Strategies

Table 6 reports macro-F1 under the same backbones, budgets, and baselines as Table 1. Accuracy counts correct predictions over all images, so a method can score well by predicting frequent or visually dominant emotions correctly while failing on the rest. Macro-F1 averages the F1 score of each class with equal weight and therefore also reflects whether less frequent emotions are recognized. This section expands the summary in Section 4.2.

Overall ranking. EMOPRUNE achieves the best average macro-F1 in all four pruning settings. Its margin over the strongest baseline is larger than under accuracy in every setting, and the gap widens as the budget tightens: on LLaVA-1.5-7B, the margin grows from 2.10 points at 10% to 4.04 points at 5%. This mirrors Table 2, where the advantage of the learned head over `[CLS]` attention is also largest under aggressive pruning. On LLaVA-1.5-7B at 10%, EMOPRUNE exceeds full-token

Table 6: Comparison of macro-F1 (%) and per-image cost on two models. Macro-F1 is reported on six datasets; Avg. is their unweighted mean. TFLOPs/img is the per-image mean. No pruning uses all visual tokens; for pruning methods, β is the drop rate (β=0.9 keeps 10% and β=0.95 keeps 5%). In the Avg. column, bold is best in each block and underline is second best. EMOPRUNE reports mean±std over three independent retrainings (data resampling and initialization).

		Function	**FI** 8	**Emotion6** 6	**ArtPhoto** 8	**EmoSet** 8	**UnbiasedEmo** 6	**WebEmo$_{25}$** 25	**Avg.**	**TFLOPs/img↓**
LLaVA-1.5-7B		No pruning	56.46	31.27	36.92	51.86	76.64	11.96	44.19	8.8
	β=0.9	SEPM	38.91	28.68	30.77	37.06	12.79	2.42	25.11	12.0
		FastV	44.42	11.90	27.87	38.27	42.75	9.89	29.18	2.2
		SparseVLM	57.05	29.46	35.50	51.03	66.32	11.51	41.81	2.7
		FasterVLM	53.53	35.46	35.18	47.64	75.30	13.22	<u>43.39</u>	1.78
		VisionZip	50.41	30.81	33.67	44.70	70.72	10.25	40.09	1.78
		PruneSID	55.92	30.60	37.59	49.34	74.00	10.86	43.05	1.78
		EMOPRUNE	54.88±0.71	38.96±0.40	37.43±1.27	51.02±1.55	78.62±0.27	12.01±0.34	**45.49**±0.69	1.78
	β=0.95	SEPM	27.70	19.23	21.87	26.80	13.14	2.07	18.47	11.6
		FastV	30.49	7.43	21.66	26.88	23.43	6.92	19.47	1.85
		SparseVLM	52.39	24.51	32.25	47.41	61.60	10.10	38.04	2.2
		FasterVLM	47.85	30.67	31.14	42.84	70.62	10.29	<u>38.90</u>	1.39
		VisionZip	43.70	21.81	28.74	38.12	61.96	8.40	33.79	1.39
		PruneSID	50.40	21.87	35.24	43.67	68.49	8.58	38.04	1.39
		EMOPRUNE	53.15±1.33	33.89±0.26	35.69±0.97	47.46±2.44	77.50±0.15	9.97±0.25	**42.94**±0.53	1.39
InternVL3-8B		No pruning	59.88	45.03	47.89	57.50	78.61	19.73	51.44	5.7
	β=0.9	SEPM	51.69	38.47	39.43	53.52	58.45	16.34	42.98	6.8
		FasterVLM	50.61	38.71	40.85	52.79	74.24	16.38	<u>45.60</u>	2.2
		VisionZip	52.77	38.63	39.57	53.07	70.15	14.14	44.73	2.2
		PruneSID	53.40	40.34	39.65	53.23	69.80	14.15	45.10	2.2
		EMOPRUNE	52.54±1.02	40.65±0.19	42.53±0.91	54.36±0.88	71.99±1.80	16.48±1.23	**46.43**±0.86	2.2
	β=0.95	SEPM	47.20	35.00	36.17	49.19	50.64	11.47	38.28	6.7
		FasterVLM	40.20	28.58	33.40	43.49	61.58	12.69	36.66	2.0
		VisionZip	48.74	35.17	32.74	47.98	61.84	9.60	<u>39.35</u>	2.0
		PruneSID	45.26	34.60	32.81	46.94	64.19	10.90	39.12	2.0
		EMOPRUNE	44.93±1.78	36.29±0.95	37.21±1.63	47.90±1.14	64.83±3.64	12.12±0.27	**40.54**±0.50	2.0

inference in average macro-F1 (45.49 ± 0.69 vs. 44.19) and on four of the six datasets, compared with three under accuracy. At 5%, it achieves the best macro-F1 on five of the six datasets.

Accuracy versus macro-F1. The ranking of the baselines changes between the two metrics. On LLaVA-1.5-7B at 5%, SparseVLM is second in average accuracy but third in macro-F1, and on InternVL3-8B at 5%, SEPM leads in accuracy but falls to fourth in macro-F1. The difference between accuracy and macro-F1 makes this explicit: on LLaVA-1.5-7B, it averages 2.10 and 2.16 points for EMOPRUNE at 10% and 5%, the smallest of all methods, compared with 2.81–4.85 points for the vision-side and SparseVLM baselines, 3.62 for No pruning, over 7 for FastV, and over 12 for SEPM.

Where the gap arises. The largest gaps concentrate on specific datasets. On UnbiasedEmo, SEPM reaches 43.80% accuracy at 10% on LLaVA-1.5-7B but only 12.79 macro-F1, and SparseVLM loses more than 11 points between the two metrics, whereas EMOPRUNE keeps the highest value under both. On WebEmo$_{25}$ with LLaVA-1.5-7B, SEPM has the highest accuracy among pruning methods at both budgets yet the lowest macro-F1, around 2, which indicates that its predictions collapse onto a few of the 25 classes. These cases show that accuracy alone can hide a collapse of predictions onto a few classes under aggressive pruning, which is why we report macro-F1 alongside accuracy.

B.2 MACRO-F1 VERSUS COMPUTE ON ALL DATASETS

Figure 7 extends Figure 4 to the remaining four datasets and shows the same overall trend: EMOPRUNE remains among the strongest methods across the compute range, and its advantage is most visible at low compute, where the baselines lose macro-F1 fastest. At 50% the methods are close to one another, and the gap widens as compute decreases. FastV declines most steeply on all four

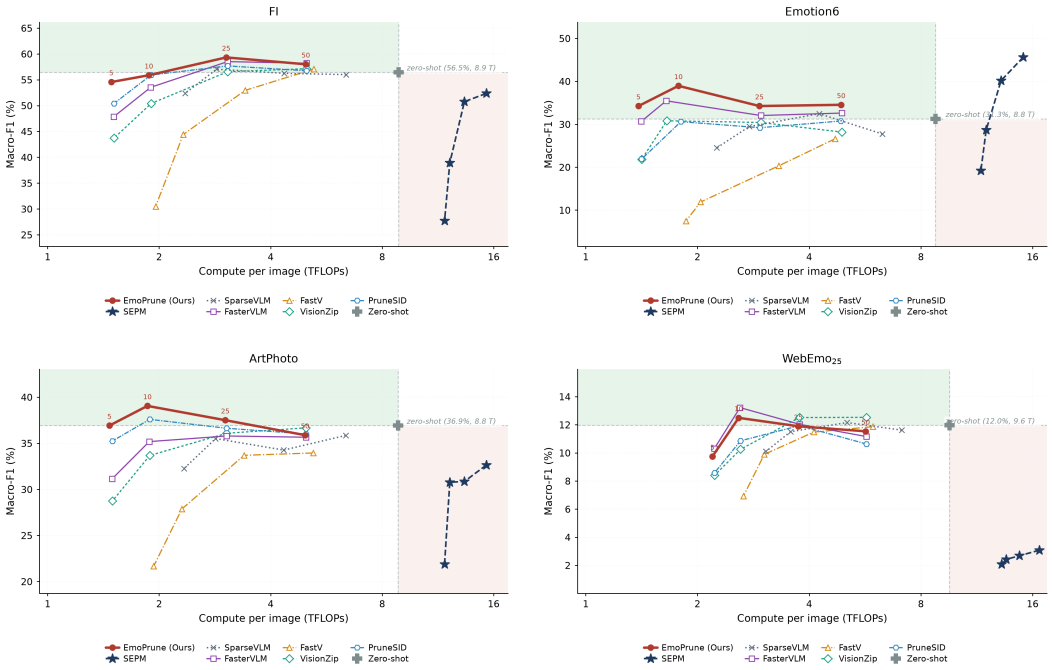


Figure 7: Macro-F1 versus per-image compute on FI, Emotion6, ArtPhoto, and WebEmo$_{25}$ with LLaVA-1.5-7B. Points on each curve correspond to keep ratios of 5%, 10%, 25%, and 50%; the cross marks the No pruning reference.

datasets. On FI, Emotion6, and ArtPhoto, EMOPRUNE stays close to or above the No pruning level at most keep ratios, and SEPM costs more than No pruning on all four datasets.

C Additional Qualitative Examples

C.1 [CLS] Attention versus Evidence Head

Selection overlap. To quantify how much the two signals agree, we compute, for each test image, the intersection-over-union (IoU) between the top-K patches ranked by [CLS] attention and by the Evidence Head, with both scores taken from the same forward pass of LLaVA-1.5-7B. As a reference, two independent random selections at keep ratio r have an expected IoU of $r/(2-r)$. Table 7 and Figure 8 report the results on the six test sets. At every keep ratio, the overlap is 2.0–3.4 times the random level, yet at the small budgets used in our main experiments it remains low: at 10%, the IoU is 0.12, so about one fifth of the patches selected by the head are also selected by [CLS]. The two signals therefore share part of the selected region, while most of their selections differ.

Table 7: IoU between the top-K selections of [CLS] attention and the Evidence Head on LLaVA-1.5-7B. Mean and range are taken over the six test sets; Random is the expected IoU of two independent random selections.

Keep ratio	Random	Mean	Range	Mean / Random
5%	0.026	0.089	0.076–0.101	3.4×
10%	0.053	0.116	0.096–0.139	2.2×
25%	0.143	0.284	0.243–0.321	2.0×
50%	0.333	0.675	0.617–0.736	2.0×

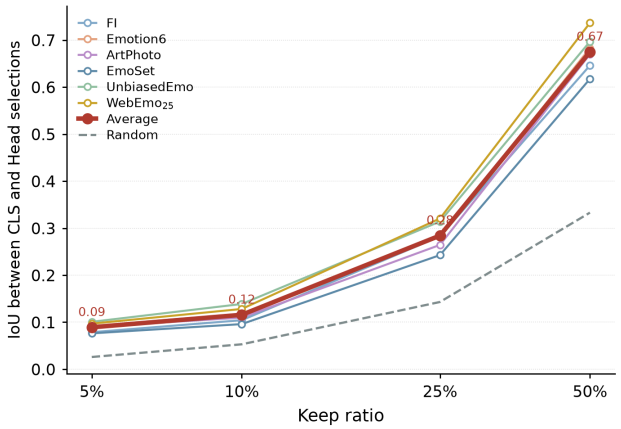


Figure 8: IoU between the top-K selections of [CLS] attention and the Evidence Head on LLaVA-1.5-7B across keep ratios. Thin lines show individual test sets, the thick line their mean, and the dashed line the expected IoU of two independent random selections.

Qualitative examples. Figures 9–12 visualize the token selections at keep ratios of 5%, 10%, 25%, and 50%, one figure per ratio with three examples each. In each example, panel (a) shows the input image with the ground-truth label; panels (b)–(d) show the patches kept by [CLS] attention, the Evidence Head, and their combination, each with its prediction; and panel (e) is a difference map in which blue marks patches kept only by [CLS], orange marks patches kept only by the head, dimmed patches are kept by both, and black patches are discarded by both.

Across budgets the two signals show consistently different preferences. [CLS] attention is concentrated on globally salient content: it keeps the most prominent subject and high-contrast surroundings, and the type of region it selects stays similar as the budget shrinks. The Evidence Head selects more local, emotion-expressive evidence, such as faces, hands, limb posture, and objects under interaction, which often lies outside the salient region. In the middle row of Figure 9, it selects patches at field level near the legs and the ball, whereas [CLS] keeps the players' upper bodies; in the top row, it selects fewer patches on the backlit subject on which [CLS] is concentrated. The dominance

of blue and orange over dimmed patches in the difference maps is consistent with the low overlap in Table 7. The combination keeps both the salient context from [CLS] and the head's local cues. In the last row of Figure 10, both signals are wrong individually, while the combination, which keeps the bright lantern together with the head's patches, predicts the correct label.

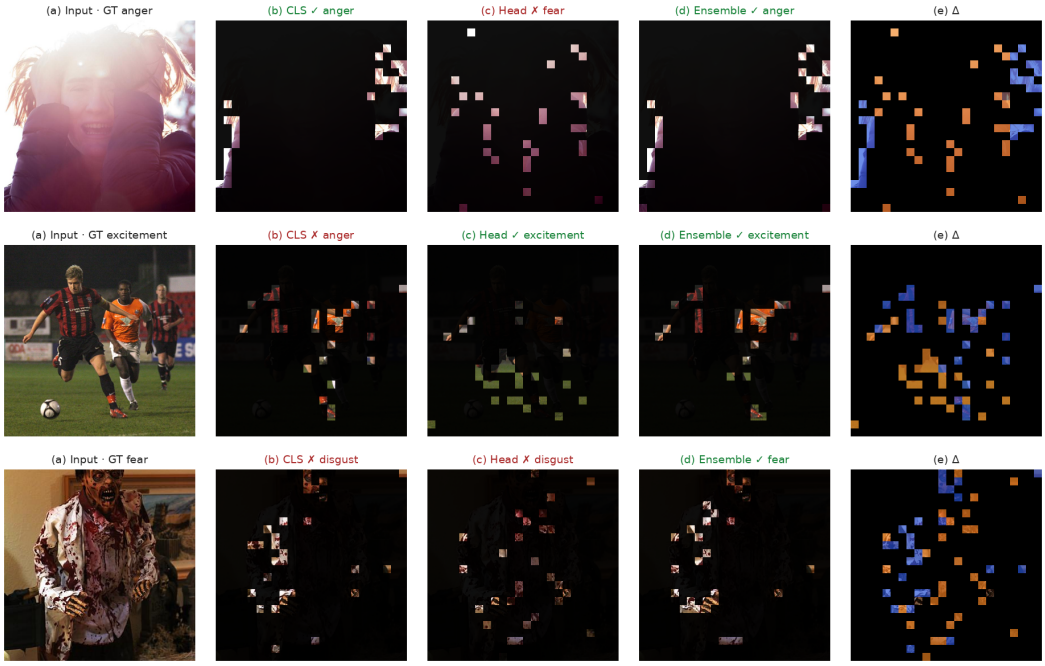


Figure 9: Token selections of [CLS] attention, the Evidence Head, and their combination at keep 5%. Rows from top have true labels anger, excitement, and fear.

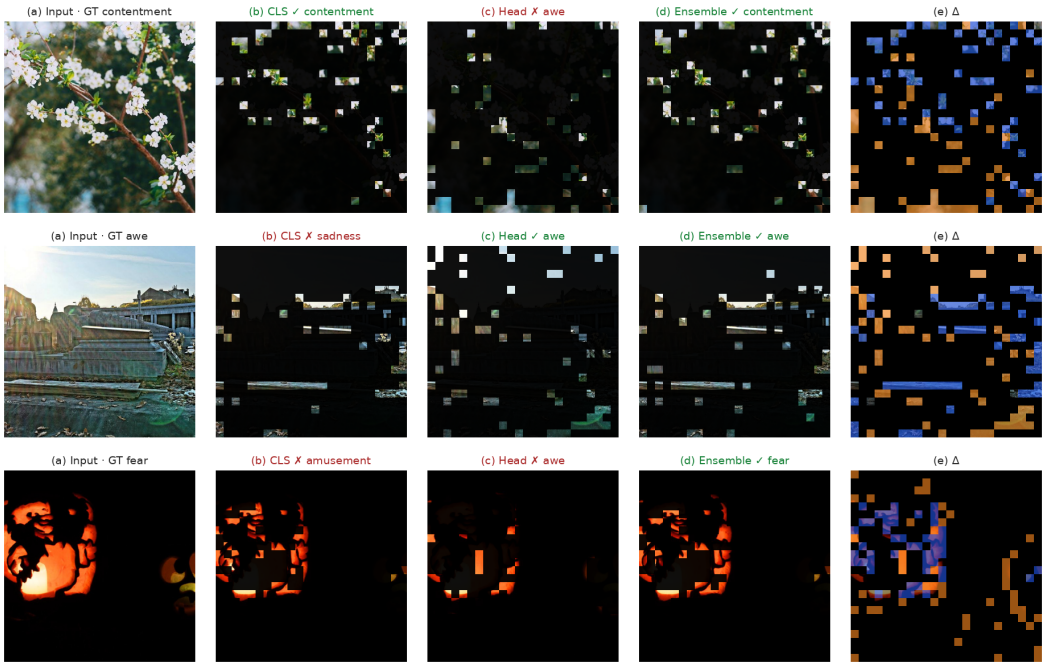


Figure 10: Token selections at keep 10%, as in Figure 9. Rows from top have true labels contentment, awe, and fear.

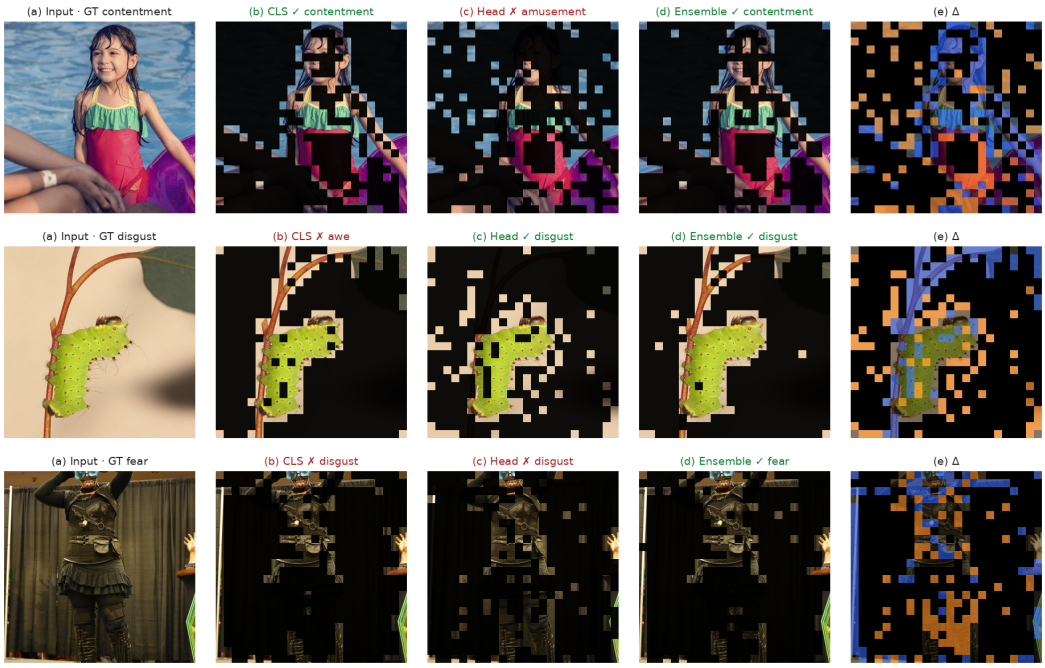


Figure 11: Token selections at keep 25%, as in Figure 9. Rows from top have true labels contentment, disgust, and fear.



Figure 12: Token selections at keep 50%, as in Figure 9. Rows from top have true labels contentment, sadness, and excitement.

D Error Analysis

We group the failure cases of the three scoring signals into three types, shown in Figures 13–15 in the same format as Figure 5: the original image with its label, followed by the tokens kept by `[CLS]` attention, the Evidence Head, and their combination, each with its prediction.

The combination overrides a correct head. In Figure 13, the head alone predicts the correct label, but the combination does not. In the top row, `[CLS]` concentrates on the brightly lit lower part of the toy, whereas the head keeps only a few patches there; the combination is dominated by this region and predicts "contentment." In the bottom row, the head keeps the bared teeth and open mouth of the tiger, and `[CLS]` keeps the surrounding fur; the combination mixes the two and predicts "fear." In both cases the two signals favor different regions and the fixed weight α does not keep enough of the head's evidence, which corresponds to the first limitation discussed in the main text.

The head selects ambiguous or background cues. In Figure 14, `[CLS]` alone is correct, while the head and the combination are not. In the top row, `[CLS]` keeps parts of the girl, whereas the head spreads its budget over the grass and the scattered figures in the background, and the prediction shifts from "amusement" to "awe." In the bottom row, the head keeps the red gloves and the contact between the boxers, which signal an intense interaction but do not by themselves separate "anger" from "excitement." Such cues are emotion-relevant but shared by neighboring high-arousal classes.

All signals fail. In Figure 15, all three signals predict the same wrong label. In the top row, the jumping pose conveys high arousal, and every selection predicts "excitement" instead of "amusement," two classes that are hard to separate from posture alone. In the bottom row, all selections keep the fire and predict "awe," while the "anger" label depends on a symbolic reading of the whole scene rather than on a local cue. These cases reflect the ambiguity of neighboring labels and the limitation of independent patch scoring for scene-level emotion noted in the main text.

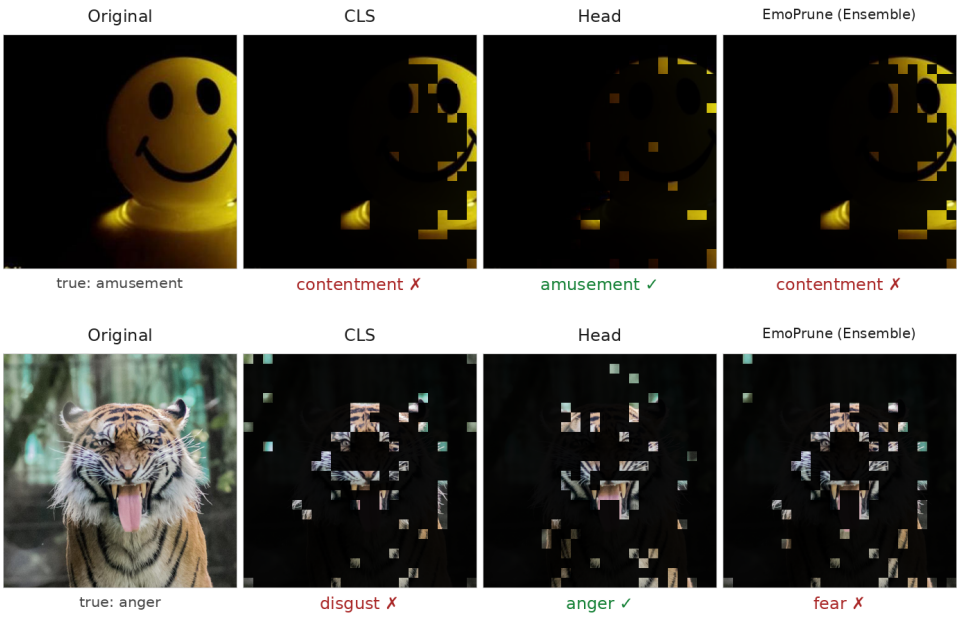


Figure 13: Failure cases where the Evidence Head is correct but the combination is not. True labels from top: amusement and anger.

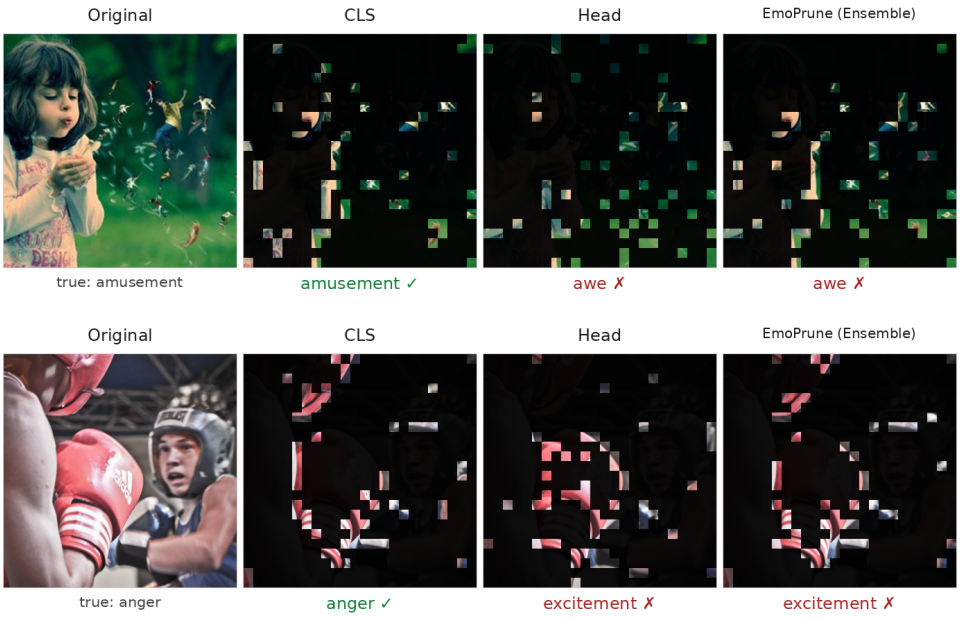


Figure 14: Failure cases where [CLS] attention is correct but the Evidence Head and the combination are not. True labels from top: amusement and anger.



Figure 15: Failure cases where all three signals predict the same wrong label. True labels from top: amusement and anger.