DEPTHROUTE: ADAPTIVE VISUAL-DEPTH SELECTION FOR MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sentiment analysis has mainly focused on how visual and textual features are fused, while paying less attention to how representations from different depths of a pretrained visual encoder should be utilized before fusion. Representations from multiple visual depths may provide complementary information, yet existing approaches typically rely on the final visual layer or aggregate multiple layers using a fixed strategy. Such choices impose a largely input-independent depth-utilization pattern, although the visual cues relevant to sentiment can vary across image–text pairs and not all information encoded at different depths is equally useful for sentiment prediction. We therefore formulate visual-depth utilization as an input-dependent, sentiment-oriented representation-selection problem and introduce DepthRoute. DepthRoute explores complementary routing configurations that vary in scope and granularity: layer-wise routing over the full visual hierarchy, routing restricted to deeper layers, and structured routing over contiguous depth groups. All variants use a shared expert pool, with routing conditioned on the paired textual representation and optimized jointly with the sentiment objective. Experiments on MVSA-Single and TumEmo with frozen SigLIP2 and CLIP visual encoders demonstrate the effectiveness of adaptive visual-depth routing over fixed final-layer representations and fixed multi-layer aggregation. An expert-enhanced final-layer control shows that additional expert capacity explains only part of the improvement. Late-restricted routing remains highly competitive, while structured full-depth routing achieves the strongest overall mean performance in our controlled comparisons. Class-wise routing analysis further reveals sentiment-dependent preferences over depth-associated routes, showing that the learned routing behavior varies with sentiment categories.

1 INTRODUCTION

Multimodal sentiment analysis aims to infer affective polarity from paired textual and visual content, such as social media posts containing text and images Niu et al. (2016); Xu & Mao (2017). Recent work has largely focused on how textual and visual information should interact, including multi-view calibration Wei et al. (2023), cross-modal feature alignment Zhang et al. (2024), dynamic routing Chen et al. (2024b), and fine-grained visual modeling Yang et al. (2021; 2022); Zhao et al. (2026); Li et al. (2025). These approaches have substantially advanced how heterogeneous sentiment cues are modeled and combined across modalities. Yet before visual information is fused with text, a more basic question remains: should multimodal sentiment models rely primarily on the final visual layer, or explicitly consider representations from multiple depths of a pretrained visual encoder?

Pretrained visual encoders naturally form a hierarchy of representations across depth: lower layers tend to preserve local appearance, intermediate layers capture more structured region- and object-level information, and higher layers increasingly encode abstract semantics Zeiler & Fergus (2014); Lin et al. (2025). Recent vision–language studies have further shown that intermediate visual representations can benefit multimodal modeling through hierarchical feature integration, multi-layer connectors, and intermediate feature injection Meng et al. (2024); Yao et al. (2024); Bai et al. (2025); Lin et al. (2025). These findings suggest that sentiment-relevant information need not be concentrated solely in the final visual layer, and that representations from multiple depths may provide complementary evidence.

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.

However, merely accessing multiple visual layers does not resolve how they should be utilized for individual inputs. Existing strategies often either rely on the final visual layer or combine multiple depths using a fixed aggregation rule. Such designs apply essentially the same depth-utilization pattern across image–text pairs, even though the visual cues relevant to sentiment can vary substantially from one pair to another. More importantly, not all information encoded across visual depths is equally useful for sentiment prediction: some layers may emphasize visual content that is less relevant to the affective cues needed for the current image–text pair. Indiscriminately aggregating multiple layers may therefore introduce redundant or sentiment-irrelevant information. This motivates us to treat visual-depth utilization as an *input-dependent, sentiment-oriented representation-selection problem*: rather than fixing visual depth in advance, can a model adaptively determine which visual representations are most useful for sentiment prediction before multimodal fusion?

To investigate this question, we introduce **DepthRoute**, a framework for adaptive visual-depth selection. We study complementary routing configurations that differ in both routing scope and routing granularity. *Layer-wise routing* selects among individual layers across the full visual hierarchy, directly exposing all encoder depths as routing candidates. A *late-restricted routing* configuration limits the candidate space to deeper layers, allowing us to examine whether adaptive selection within late representations is already sufficient. Finally, *structured routing* organizes neighboring layers into contiguous depth groups and performs routing over these structured regions. Across all configurations, routing is conditioned on the paired textual representation and optimized jointly with the downstream sentiment objective through a shared expert pool, enabling input-dependent visual-depth selection that is aligned with the sentiment task.

This formulation enables us to study several questions within a common framework. First, does adaptive visual-depth selection improve over the standard use of the final visual layer or fixed multi-layer aggregation? Second, is adaptive selection within deeper layers sufficient, or does access to a broader range of visual depths remain useful? Third, how does structuring visual depths into contiguous groups compare with direct routing over individual layers? These comparisons allow us to study visual-depth utilization itself while controlling for alternative explanations such as increased expert capacity.

We evaluate DepthRoute on **MVSA-Single** and **TumEmo** using frozen **SigLIP2** and **CLIP** visual encoders. Across both datasets and visual backbones, late-restricted and structured group-wise routing consistently improve over final-layer representations and fixed multi-layer aggregation, while full-hierarchy layer-wise routing is more backbone-sensitive. Late-restricted routing remains highly competitive, showing that deeper visual representations provide strong candidates for adaptive selection, while structured full-depth routing achieves the strongest overall mean performance in our controlled comparisons. An expert-enhanced final-layer control shows that additional expert capacity contributes to performance, but does not fully account for the gains of adaptive visual-depth routing. Moreover, class-wise routing analysis reveals sentiment-dependent preferences over depth-associated routes, providing further evidence that the learned routing behavior is shaped by the downstream sentiment objective.

Our main contributions are summarized as follows:

- We formulate visual-depth utilization as an *input-dependent, sentiment-oriented representation-selection problem* for multimodal sentiment analysis, complementing the conventional focus on how textual and visual information is fused.
- We introduce **DepthRoute**, a unified framework for adaptive visual-depth routing, and study complementary routing configurations over individual layers, late-depth subsets, and contiguous depth groups using a shared expert pool.
- We conduct controlled evaluations on MVSA-Single and TumEmo with SigLIP2 and CLIP visual encoders. DepthRoute-Late and DepthRoute-Group consistently improve over final-layer representations and fixed multi-layer aggregation, while expert-enhanced controls and class-wise routing analyses further investigate the effects of expert capacity, routing scope, routing granularity, and sentiment-dependent routing behavior.

2 Related Work

Multimodal sentiment analysis. Multimodal sentiment analysis has primarily focused on modeling interactions between textual and visual information, including cross-modal alignment, dynamic fusion, and fine-grained visual modeling Wei et al. (2023); Zhang et al. (2024); Chen et al. (2024b); Yang et al. (2021; 2022); Zhao et al. (2026); Li et al. (2025). These methods mainly study how textual and visual information should be combined. In contrast, we focus on a complementary question: *how should representations from different visual encoder depths be utilized for multimodal sentiment analysis?*

Multi-layer visual representations. Intermediate visual representations have been increasingly explored in vision–language models through hierarchical feature integration, multi-layer connectors, and intermediate feature injection Meng et al. (2024); Yao et al. (2024); Bai et al. (2025); Lin et al. (2025). These approaches show that useful visual information is distributed across encoder depth, but typically rely on predefined or fixed mechanisms for combining multiple layers. Our work instead treats visual-depth utilization as an input-dependent representation-selection problem.

Mixture-of-experts and adaptive routing. Mixture-of-experts architectures provide a flexible mechanism for conditional computation and expert specialization, and have been widely explored in both unimodal and multimodal learning Shazeer et al. (2017); Fedus et al. (2022); Lepikhin et al. (2020); Riquelme et al. (2021); Du et al. (2022); Chen et al. (2025; 2024b). Recent multimodal variants further use routing to adapt feature interaction or expert utilization across inputs. Different from these works, DepthRoute applies adaptive routing explicitly over representations from different visual encoder depths and studies how routing scope and granularity affect sentiment-oriented visual-depth selection.

3 Method

3.1 Overview

Given an image–text pair (V, T), multimodal sentiment analysis aims to predict a sentiment label $y \in \mathcal{Y}$, where $|\mathcal{Y}| = C$. We use BERTweet to encode the text into $\mathbf{h}^t \in \mathbb{R}^d$. For the image, a frozen pretrained visual encoder exposes hidden states from L layers, $\mathbf{H}^v = \{\mathbf{h}_1^v, \ldots, \mathbf{h}_L^v\}$, where $\mathbf{h}_l^v \in \mathbb{R}^{N \times d_v}$ denotes the token-level representation from visual layer l.

Rather than relying only on the final visual layer or applying a fixed aggregation rule across multiple layers, DepthRoute treats visual depth as an adaptive representation-selection space. As shown in Fig. 2, DepthRoute consists of two core modules: **Visual-Depth Candidate Construction** and **Depth–Expert Routing**. Section 3.3 describes how visual representations from different encoder depths are organized into candidate sets with different routing scopes and granularities. Section 3.4 then introduces the shared Depth-Expert Routing mechanism, which adaptively routes these candidates through a shared expert pool conditioned on the paired textual representation. The resulting routed visual representation is combined with the textual representation for sentiment prediction.

3.2 Visual-Depth Selection Strategies

Figure 1 illustrates three representative ways of utilizing representations from a pretrained visual encoder: final-layer representation, fixed multi-layer aggregation, and adaptive visual-depth selection.

Final-layer representation. A common strategy is to use only the output of the final visual encoder layer. Given the multi-layer hidden states $\mathbf{H}^v = \{\mathbf{h}_1^v, \ldots, \mathbf{h}_L^v\}$, the resulting visual representation is

$$\mathbf{v}_{\text{final}} = \text{Pool}(\mathbf{h}_L^v), \tag{1}$$

where $\text{Pool}(\cdot)$ converts token-level hidden states into a compact visual representation. This strategy applies the same final-depth choice to every input.

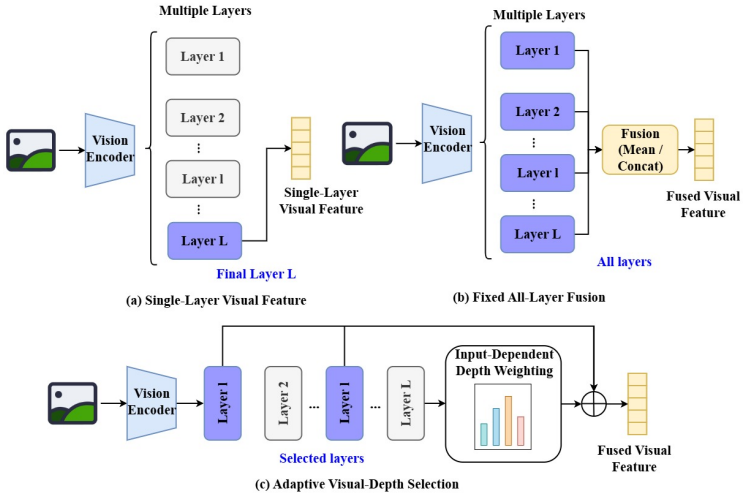


Figure 1: **Comparison of visual-depth utilization strategies.** (a) **Final-layer representation** uses only the representation from the last visual encoder layer. (b) **Fixed all-layer aggregation** combines representations from multiple visual depths using the same aggregation strategy for all inputs. (c) **Adaptive visual-depth selection** allows the contribution of different visual depths to vary across inputs, enabling input-dependent selection of visual representations.

Fixed multi-layer aggregation. A direct way to exploit representations beyond the final layer is to aggregate features from multiple visual depths. Given a selected layer set $\mathcal{S}$, we write this generally as

$$\mathbf{v}_{\text{fixed}} = \text{Agg}\left(\{\text{Pool}(\mathbf{h}_l^v)\}_{l\in\mathcal{S}}\right), \tag{2}$$

where $\text{Agg}(\cdot)$ can represent averaging, concatenation, or globally learned scalar weighting. Although this strategy incorporates information from multiple depths, the same aggregation rule is shared across all image–text pairs.

Adaptive visual-depth selection. DepthRoute instead treats the visual hierarchy as an adaptive selection space. Rather than fixing one visual depth or applying the same multi-layer aggregation to every input, it constructs multiple visual-depth candidates and learns how they should contribute to each image–text pair. This enables visual-depth utilization to vary with the current input and the downstream sentiment objective. We next describe how these candidates are constructed and routed within DepthRoute.

3.3 Visual-Depth Candidate Construction

Given the multi-layer visual representations $\mathbf{H}^v = \{\mathbf{h}_1^v, \ldots, \mathbf{h}_L^v\}$, DepthRoute constructs a set of visual-depth candidates $\mathcal{D} = \{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$. Each candidate represents either an individual visual layer or a contiguous region of the visual hierarchy. By changing the candidate set while keeping the subsequent routing mechanism unchanged, we study visual-depth selection under different routing scopes and granularities.

Layer-wise candidates. The most direct configuration treats every visual layer as an individual candidate. For the l-th visual layer, we obtain

$$\mathbf{v}_l = \text{MeanPool}(\mathbf{h}_l^v), \qquad l = 1, \ldots, L, \tag{3}$$

which gives the candidate set

$$\mathcal{D}_{\text{layer}} = \{\mathbf{v}_1, \ldots, \mathbf{v}_L\}. \tag{4}$$

This configuration exposes the full visual hierarchy to the router and allows the contribution of each layer to vary across inputs.

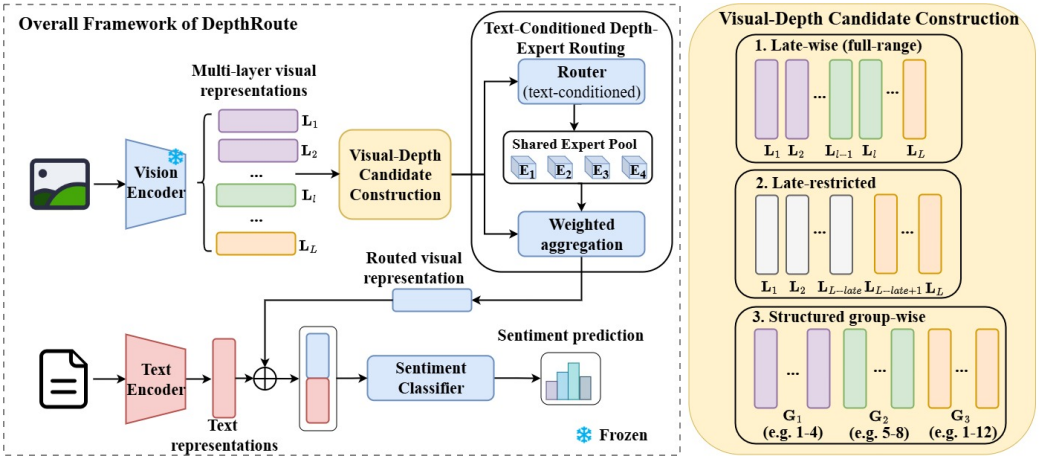


Figure 2: **Overview of DepthRoute.** The main pipeline extracts multi-depth representations from a frozen visual encoder, constructs visual-depth candidates, and adaptively routes them through a shared expert pool conditioned on the paired textual representation. The candidate-construction module is instantiated with three complementary configurations: layer-wise candidates over the full visual hierarchy, late-restricted candidates over deeper layers, and structured candidates formed from contiguous depth groups. All configurations share the same routing mechanism.

Late-restricted candidates. To examine whether adaptive selection within deeper visual representations is already sufficient, we restrict the candidate set to the late visual layers:

$$\mathcal{D}_{\text{late}} = \{\mathbf{v}_l \mid l \in \mathcal{L}_{late}\}, \tag{5}$$

where $\mathcal{L}_{late}$ denotes the set of late visual layers. This configuration retains layer-wise routing while restricting the routing scope to the deeper part of the visual encoder.

Structured group-wise candidates. We further organize neighboring visual layers into contiguous depth groups $\mathcal{G} = \{G_1, \ldots, G_{N_g}\}$. For each group G_g, the token-level hidden states of its constituent layers are concatenated along the feature dimension, projected into the shared hidden space, and mean-pooled:

$$\mathbf{v}_g = \text{MeanPool}\left(\phi_g\left([\mathbf{h}_l^v]_{l \in G_g}\right)\right), \tag{6}$$

where $\phi_g(\cdot)$ denotes a group-specific projection.

3.4 Depth-Expert Routing

Given the visual-depth candidate set $\mathcal{D} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{N_c}\}$, DepthRoute routes each candidate through a shared pool of N_e experts. Let $\{F_e(\cdot)\}_{e=1}^{N_e}$ denote the shared experts. For candidate $\mathbf{v}_i$, expert e produces

$$\mathbf{o}_{i,e} = F_e(\mathbf{v}_i). \tag{7}$$

The routing decision is conditioned on the paired textual representation $\mathbf{h}^t$. For each candidate–expert pair, the router predicts a routing logit $r_{i,e}$ together with a route-specific log-scale $\ell_{i,e}$:

$$r_{i,e} = f_{\text{route}}(\mathbf{v}_i, \mathbf{h}^t)_e, \qquad \ell_{i,e} = f_{\text{scale}}(\mathbf{v}_i, \mathbf{h}^t)_e. \tag{8}$$

The log-scale provides an uncertainty-aware adjustment to the routing preference. We define the adjusted routing score as

$$s_{i,e} = r_{i,e} - \lambda_{\text{lap}} \ell_{i,e}, \tag{9}$$

where λ_{lap} controls the contribution of the route-scale adjustment.

The adjusted scores are jointly normalized over all candidate–expert pairs:

$$\beta_{i,e} = \frac{\exp(s_{i,e}/\tau)}{\sum_{i'=1}^{N_c} \sum_{e'=1}^{N_e} \exp(s_{i',e'}/\tau)}, \tag{10}$$

where τ is the routing temperature.

The final routed visual representation is obtained by weighted aggregation:

$$\mathbf{h}^v = \mathrm{LN}\left(\sum_{i=1}^{N_c}\sum_{e=1}^{N_e}\beta_{i,e}\mathbf{o}_{i,e}\right). \tag{11}$$

Since the routing scores depend on both the visual-depth candidates and the paired text, DepthRoute can adapt its visual-depth utilization across different image–text pairs. The routing module, route-scale predictor, and shared experts are trained jointly with the downstream sentiment objective.

3.5 Sentiment Prediction and Training Objective

The routed visual representation $\mathbf{h}^v$ is concatenated with the textual representation $\mathbf{h}^t$ and fed into an MLP classifier:

$$\hat{\mathbf{y}} = \mathrm{Softmax}\left(\mathrm{MLP}_c\left([\mathbf{h}^t;\mathbf{h}^v]\right)\right), \tag{12}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^C$ denotes the predicted probability distribution over C sentiment classes.

The primary training objective is the cross-entropy loss:

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{B}\sum_{n=1}^{B}\sum_{c=1}^{C} y_{n,c}\log\hat{y}_{n,c}, \tag{13}$$

where B denotes the batch size and $y_{n,c}$ is the ground-truth indicator for class c.

We further regularize the magnitude of the learned route log-scales:

$$\mathcal{L}_{\mathrm{scale}} = \frac{1}{N_c N_e}\sum_{i=1}^{N_c}\sum_{e=1}^{N_e}|\ell_{i,e}|, \tag{14}$$

where N_c and N_e denote the numbers of visual-depth candidates and experts, respectively.

The final training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{scale}}\mathcal{L}_{\mathrm{scale}}, \tag{15}$$

where λ_{scale} controls the strength of the scale regularization.

All trainable components, including the text encoder, visual-depth candidate projections, router, scale head, shared expert pool, and sentiment classifier, are optimized jointly under this objective, while the pretrained visual encoder remains frozen.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate DepthRoute on MVSA-Single Niu et al. (2016) and TumEmo Yang et al. (2020). MVSA-Single contains 3,611/450/450 train/validation/test samples, while TumEmo contains 156,204/19,525/19,536. We report Accuracy and Weighted-F1, and use validation Weighted-F1 for model selection.

Comparison methods. We compare DepthRoute with representative published multimodal sentiment methods and controlled visual-depth baselines. The controlled baselines are *Final Layer* and fixed *All-Layer Fusion*. We evaluate three DepthRoute variants: *Layer* for full-hierarchy layer-wise routing, *Late* for late-layer routing, and *Group* for contiguous group-wise routing.

Implementation details. We evaluate all DepthRoute variants with two distinct pretrained visual backbones, SigLIP2-base-patch16-224 Tschannen et al. (2025) and CLIP ViT-B/16 Radford et al. (2021), to assess whether the proposed visual-depth routing generalizes across encoder families. Both visual backbones are frozen during training, whereas the text encoder and DepthRoute modules are trainable. DepthRoute uses four shared experts and a 4–4–4 partition for group-wise routing. We set $\tau = 1.0$ on MVSA-Single, $\tau = 0.8$ on TumEmo, and $\lambda_{\mathrm{scale}} = 10^{-4}$. Models are trained with AdamW, and controlled results are averaged over seeds 199, 42, and 2025.

Table 1: Comparison on MVSA-Single and TumEmo. Published results are taken from the corresponding papers, while controlled baselines and DepthRoute variants are reported as mean±std over three seeds.

	MVSA-Single		TumEmo	
Method	Acc.	W-F1	Acc.	W-F1
Recent methods				
D2R Chen et al. (2024b)	0.7667	0.7559	–	–
CF-MSA Chen et al. (2024a)	0.7617	0.7454	–	–
DIB Huang et al. (2025)	0.7605	0.7520	–	–
Global Context Liu & Qian (2025)	0.7511	0.7323	0.6918	0.6898
MapleNet Yang et al. (2025)	0.7010	0.6869	0.6633	0.6639
SigLIP2				
Final	0.7546±0.0067	0.7468±0.0072	0.6924±0.0023	0.6921±0.0021
Final + Expert	0.7674±0.0082	0.7530±0.0097	0.7132±0.0013	0.7129±0.0016
All-Layer	0.7659±0.0053	0.7516±0.0064	0.7065±0.0018	0.7069±0.0017
DepthRoute-Layer	0.7778±0.0083	0.7651±0.0096	0.7091±0.0028	0.7079±0.0042
DepthRoute-Late	0.7681±0.0028	0.7590±0.0051	0.7187±0.0038	0.7189±0.0033
DepthRoute-Group	**0.7830±0.0071**	**0.7794±0.0073**	**0.7190±0.0038**	**0.7193±0.0035**
CLIP-B/16				
Final	0.7459±0.0136	0.7339±0.0124	0.6756±0.0006	0.6753±0.0005
All-Layer	0.7585±0.0055	0.7498±0.0065	0.7050±0.0011	0.7051±0.0012
DepthRoute-Layer	0.7674±0.0010	0.7604±0.0049	0.7045±0.0029	0.7047±0.0021
DepthRoute-Late	0.7681±0.0169	0.7637±0.0187	0.7096±0.0011	0.7102±0.0012
DepthRoute-Group	**0.7704±0.0124**	**0.7654±0.0108**	**0.7186±0.0052**	**0.7190±0.0053**

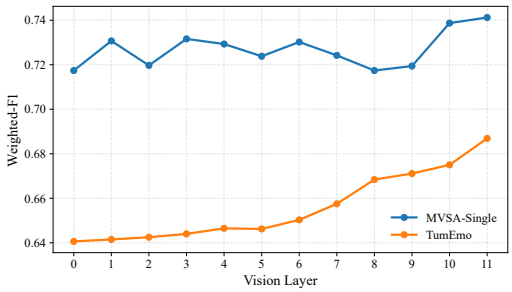


Figure 3: Layer-wise Weighted-F1 obtained by independently using each frozen SigLIP2 visual layer on MVSA-Single and TumEmo under seed 199.

4.2 Main Results

Table 1 compares DepthRoute with recent methods and controlled visual-depth baselines. DepthRoute-Group achieves the strongest controlled performance across both datasets and backbones. On MVSA-Single, it reaches 0.7794/0.7654 Weighted-F1 with SigLIP2/CLIP, compared with 0.7516/0.7498 for fixed all-layer fusion. On TumEmo, it obtains 0.7193/0.7190.

Under SigLIP2, Final+Expert improves over the plain Final baseline but remains below DepthRoute-Group on both datasets, indicating that expert capacity alone does not fully explain the gains. DepthRoute-Late and DepthRoute-Group consistently improve over the fixed baselines, while full-hierarchy layer-wise routing is more backbone-sensitive and slightly underperforms all-layer fusion on TumEmo with CLIP.

4.3 Visual-Depth Selection Analysis

To better understand why adaptive visual-depth routing is effective, we analyze the contribution of individual visual layers, fixed all-layer fusion, and different routing scopes and granularities.

Table 2: **Ablation study on MVSA-Single.** Results are reported as mean±std over three random seeds.

Method	Acc.	M-F1	W-F1
w/o Text Conditioning	0.7748±0.0013	0.6694±0.0163	0.7720±0.0051
w/o Log-Scale Adjustment	0.7771±0.0025	0.6388±0.0245	0.7636±0.0076
w/o Shared Expert Pool	0.7793±0.0090	0.6260±0.0201	0.7623±0.0110
DepthRoute-Group	**0.7830±0.0071**	**0.6789±0.0093**	**0.7794±0.0073**

Single-layer probing. Figure 3 reports the Weighted-F1 obtained by independently using each visual encoder layer. On MVSA-Single, performance varies non-monotonically with depth, with several intermediate layers remaining competitive with the final layer. In contrast, TumEmo shows a clearer preference for deeper representations, with performance improving substantially toward the final layers. These results indicate that sentiment-relevant information is distributed across visual depth and that the preferred depth pattern is dataset dependent.

Fixed all-layer fusion. The main results show that fixed all-layer fusion generally improves over using only the final layer, confirming that intermediate representations provide complementary information. However, fixed fusion applies the same depth-utilization strategy to every input. This limits its ability to adapt to image-text pairs whose sentiment-relevant visual cues may arise from different parts of the visual hierarchy.

Routing scope and granularity. DepthRoute-Layer exposes individual layers across the full hierarchy, DepthRoute-Late restricts routing to deeper layers, and DepthRoute-Group routes over contiguous depth groups. DepthRoute-Late remains consistently strong, while DepthRoute-Group achieves the best overall controlled performance across datasets and backbones. In contrast, full-hierarchy layer-wise routing is more sensitive to the visual backbone. These results suggest that adaptive visual-depth selection is beneficial, while the appropriate routing scope and granularity depend on the structure of the visual representation space.

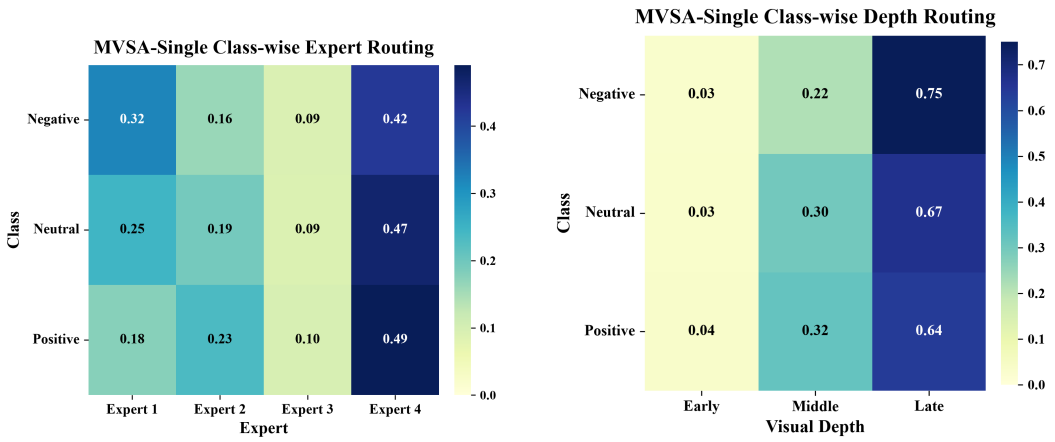
4.4 Ablation Study

We conduct controlled ablations on MVSA-Single to examine the contribution of text conditioning, the shared expert pool, and the route-specific log-scale adjustment while keeping the structured visual-depth candidate construction fixed.

Removing text conditioning reduces Weighted-F1 from 0.7794 to 0.7720, showing that routing benefits from the paired textual context. Removing the log-scale adjustment further decreases Weighted-F1 to 0.7636, indicating that route-specific uncertainty modulation contributes to more effective routing. Removing the shared expert pool results in 0.7623 Weighted-F1, showing that shared expert transformations are also important. Overall, all three components contribute to the final performance, with the largest degradations observed when the log-scale adjustment or shared expert pool is removed.

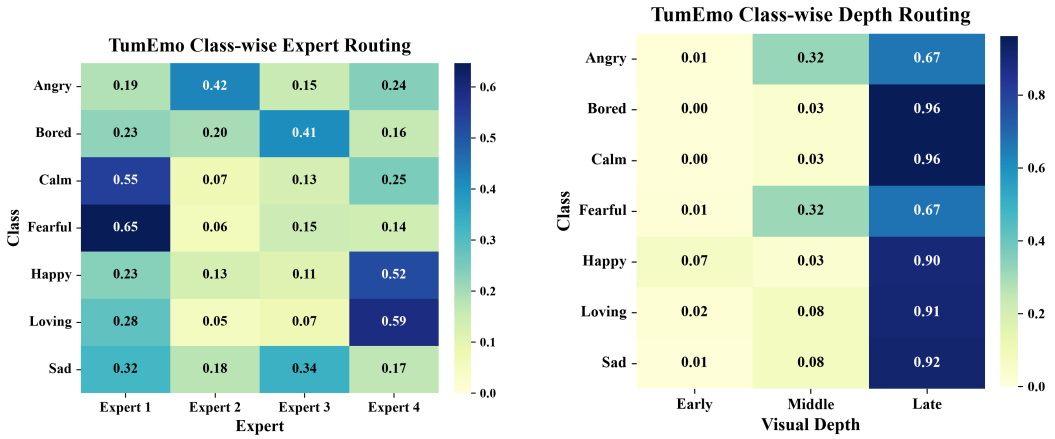
4.5 Routing Analysis

We visualize the learned routing distributions of **DepthRoute-Group** on MVSA-Single and TumEmo in Fig. 4. Different sentiment categories exhibit distinct preferences over both visual-depth groups and shared experts, showing that the learned routing behavior varies with sentiment semantics. Although late visual representations dominate overall, the relative contributions of early and middle depths differ across categories, while expert utilization also changes substantially. These class-dependent routing patterns show that DepthRoute-Group learns visual-depth and expert selection that varies across sentiment categories, rather than applying a fixed routing distribution to all inputs.



(a) MVSA-Single: expert routing.

(b) MVSA-Single: depth routing.



(c) TumEmo: expert routing.

(d) TumEmo: depth routing.

Figure 4: **Class-wise routing behavior of DepthRoute-Group on MVSA-Single and TumEmo.** Expert-routing distributions are obtained by aggregating routing probabilities over visual-depth groups, while depth-routing distributions are obtained by aggregating over experts.

5 Conclusion

We formulate visual-depth utilization in multimodal sentiment analysis as an input-dependent, sentiment-oriented representation-selection problem and introduce **DepthRoute** for adaptive routing across visual encoder depths. We study layer-wise, late-restricted, and structured group-wise routing within a shared text-conditioned expert framework. Experiments on MVSA-Single and TumEmo with SigLIP2 and CLIP show that adaptive visual-depth routing is effective, with late-restricted and structured group-wise routing consistently improving over the fixed baselines. Ablation and routing analyses further demonstrate the importance of text-conditioned routing and shared experts, while revealing class-dependent preferences over visual depths and experts that support the sentiment-driven nature of the learned routing behavior. Overall, these results highlight adaptive visual-depth selection as a complementary modeling dimension for multimodal sentiment analysis beyond conventional multimodal fusion.

AI use statement

Generative AI tools were used to assist with polishing the writing of this paper, including improving clarity, readability, and organization. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

Reproducibility Statement

We provide detailed descriptions of the proposed DepthRoute architecture, visual-depth candidate construction, routing mechanism, training objective, datasets, evaluation metrics, and implementation settings in the main paper. Controlled experiments are reported as mean and standard deviation over three random seeds. We specify the pretrained visual backbones, expert configuration, depth-group partition, routing hyperparameters, optimizer, and model-selection criterion used in our experiments. To facilitate reproducibility, the supplementary material includes the source code, experiment configurations, and additional implementation details needed to reproduce the reported results.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Fuhai Chen, Pengpeng Huang, Xuri Ge, Jie Huang, and Zishuo Bao. Multimodal sentiment analysis based on causal reasoning. *arXiv preprint arXiv:2412.07292*, 2024a.

Kezhou Chen, Shuo Wang, Huixia Ben, Shengeng Tang, and Yanbin Hao. Mixture of multimodal adapters for sentiment analysis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1822–1833, 2025.

Yifan Chen, Kuntao Li, Weixing Mai, Qiaofeng Wu, Yun Xue, and Fenghuan Li. D2r: Dual-branch dynamic routing network for multimodal sentiment detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3536–3547, 2024b.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Huiting Huang, Tieliang Gong, Kai He, Jialun Wu, Erik Cambria, and Mengling Feng. Robust multimodal sentiment analysis via double information bottleneck. *Information Fusion*, pp. 103964, 2025.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, and Dongmei Jiang. Alignmamba: Enhancing multimodal mamba with local and global cross-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24774–24784, 2025.

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multi-layer visual feature fusion in multimodal llms: Methods, analysis, and best practices. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4156–4166, 2025.

Peng Liu and Wenhua Qian. Leveraging global context for improved image–text sentiment analysis. *IEEE Transactions on Computational Social Systems*, 2025.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In *International conference on multimedia modeling*, pp. 15–27. Springer, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in neural information processing systems*, 34:8583–8595, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pp. 5240–5252, 2023.

Nan Xu and Wenji Mao. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 2399–2402, 2017.

Hao Yang, Yanyan Zhao, and Bing Qin. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 3324–3335, 2022.

Kaixiang Yang, Yifan Luo, Zongyan Zhang, CL Philip Chen, and Tong Zhang. Multimodal affect perception with large language model enhancement network. *IEEE Transactions on Affective Computing*, 2025.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026, 2020.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 328–339, 2021.

Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for mllms. *Advances in Neural Information Processing Systems*, 37:33108–33140, 2024.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Ming Zhang, Ke Chang, and Yunfang Wu. Multi-modal semantic understanding with contrastive cross-modal feature alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 11934–11943, 2024.

Xianbing Zhao, Shengzun Yang, and Buzhou Tang. Ccaf: Coarse-to-fine cross-modal alignment and fusion for multimodal sentiment analysis. In *Proceedings of the ACM Web Conference 2026*, pp. 7421–7430, 2026.