VERTICAL MEETS HORIZONTAL: DISENTANGLING AFFECTIVE GEOMETRY FOR PLUG-AND-PLAY STRUCTURED PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sentiment analysis has made substantial progress in learning expressive cross-modal representations, yet final prediction is still commonly performed by flat classifiers that treat sentiment categories as independent targets. Such representations implicitly form a task-specific affective space whose inter-class organization defines a learned affective geometry that may remain underexploited during prediction. To address this limitation, Vertical-Meets-Horizontal Prediction (VMH-P) is introduced as a plug-and-play structured prediction framework that exploits learned affective geometry from fully trained multimodal models without modifying their representation space. VMH-P aligns available multimodal representations into a shared evidence space and disentangles the resulting affective geometry along two complementary directions. Vertically, latent affective regions are induced from class-center geometry and embedded into a Lorentz hierarchy, where path-wise routing and leaf-level discrimination jointly produce hierarchical class evidence. Horizontally, centered class directions establish supportive and contrasting relations whose strengths are adaptively refined to enable inter-class evidence propagation. The resulting structured evidence is aggregated across sources and introduced into the host logits through a confidence-gated residual correction. Experiments on three multimodal affective datasets with three heterogeneous host models show consistent improvements across all nine host–dataset combinations, with average gains of 3.84% and 4.12% in ACC and F1, respectively. The code is publicly available at https://anonymous.4open.science/r/VMH-P-5692.

1 INTRODUCTION

Multimodal sentiment analysis (MSA) aims to infer affective states from heterogeneous cues conveyed by text, images, and their interactions (Cheema et al., 2021). Recent advances in multimodal learning have improved cross-modal alignment, interaction, and fusion, yielding increasingly expressive affective representations (Zhang et al., 2024; Wang et al., 2025b). These representations implicitly form a task-specific affective space, in which samples associated with different sentiment categories occupy different regions according to the representations learned by a particular model. However, the prediction stage remains dominated by conventional flat classifiers that treat sentiment categories as independent targets (Wu et al., 2024; 2026). Such classifiers separate individual classes but overlook proximity, grouping, and directional relations in the learned affective space, leaving these inter-class cues underexploited during prediction.

The affective space, however, is not inherently flat. Emotion categories exhibit continuous transitions, semantic proximity, and opposition (Cowen & Keltner, 2017; Semeraro et al., 2021), revealing two complementary forms of organization overlooked by flat prediction: hierarchical grouping across levels of granularity and relational interactions among categories at the same level. Vertically, nearby categories may form latent groups, yielding a coarse-to-fine organization from coarse sentiment polarity tendencies to fine-grained emotion categories; for example, happy and love may share one region, while sad and angry occupy another. From a horizontal perspective, categories can also exhibit different relative directions with respect to the overall affective space. Evidence favoring happy, for instance, may provide supportive information for a geometrically aligned cat-

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.

egory such as love, while providing contrasting evidence against a category such as angry. The vertical perspective therefore captures organization across different levels of granularity, whereas the horizontal perspective characterizes interactions among categories at the same decision level. Explicitly exploiting both forms of organization calls for a structured prediction mechanism beyond conventional flat classification.

Encouragingly, several recent studies have recognized the importance of inter-class relations and explored dedicated label modeling (Yin et al., 2026), fixed templates (Chen et al., 2020), or additional structural supervision (Chatterjee et al., 2021). These approaches show the value of structured label prediction, but typically introduce class structure via externally defined relations or dedicated structural modeling. By contrast, the learned affective geometry emerging from a trained host model remains implicit in continuous representations and is not directly available for structured decision reasoning. Consequently, the prediction stage still lacks a mechanism to disentangle this host-induced geometry into explicit class-wise structure, leaving useful inter-class cues underexploited.

To address this problem, Vertical-Meets-Horizontal Prediction (VMH-P) is proposed as a plug-and-play structured prediction framework that disentangles learned affective geometry into complementary vertical and horizontal structures for class-wise reasoning. VMH-P exploits this geometry without altering the representation space learned by the host model. Given cached textual, visual, and fused representations from a frozen multimodal model, VMH-P first aligns heterogeneous evidence into a shared space and derives class centers to capture their relative organization. On this basis, a vertical hyperbolic hierarchy induces latent affective regions from class-center geometry and embeds the resulting coarse-to-fine topology in Lorentz space, where path-wise routing and leaf-level discrimination jointly produce hierarchical class evidence. Complementarily, a horizontal affective relation field derives signed supportive and contrasting relations from centered class directions, then adaptively refines their strengths under geometry constraints to propagate evidence across sentiment categories. The resulting structured evidence is aggregated across sources and introduced as a confidence-gated residual correction to the host prediction. In this way, VMH-P preserves the learned representation space while restructuring prediction through complementary vertical and horizontal reasoning. The main contributions are as follows:

- This work proposes VMH-P, a backbone-agnostic plug-and-play structured prediction framework for MSA that exploits learned affective geometry without modifying or jointly retraining the host backbone.
- A vertical hyperbolic reasoning mechanism is developed to derive a coarse-to-fine hierarchy from the relative arrangement of learned class representations, supporting hierarchical routing from latent affective regions to fine-grained sentiment classes.
- VMH-P proposes a horizontal relation reasoning mechanism that characterizes supportive and contrasting interactions from relative class directions in the affective space, enabling geometry-guided evidence propagation across sentiment categories.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis integrates textual and visual cues for affect recognition, with early methods relying mainly on concatenation, attention, or cross-modal interaction (Yu & Jiang, 2019; Lin et al., 2022). Recent methods strengthen multimodal representations through contrastive alignment, adaptive fusion, and dynamic interaction. CLMLF (Li et al., 2022) combines multi-layer fusion with contrastive learning, D2R (Chen et al., 2024) adopts adaptive interaction routing, and SPP-SCL (Wu & Li, 2026) balances intra- and inter-modal representations. Other studies improve robustness through cross-modal translation (Zhang et al., 2024), feature recovery (Wu et al., 2024), semantic completion (Wang et al., 2025b), or sentiment-aware attention (Wu et al., 2026), while auxiliary cues such as color attributes (An & Wan Zainon, 2023), salient objects (Yu et al., 2023), and generated descriptions (Wang et al., 2024) further enrich affective representations.

Despite these advances, existing MSA methods mainly structure modality- or feature-level representations (Lin et al., 2022; Xu et al., 2025), whereas VMH-P targets prediction by deriving class-wise structure from learned multimodal representations.

2.2 Structured Label Modeling

Structured classification models dependencies among output categories through predefined hierarchies, learned label embeddings, or task-specific constraints. Hyperbolic geometry naturally accommodates hierarchical organization (Nickel & Kiela, 2017) and has been used to embed predefined hierarchies (Chen et al., 2020), infer latent class structures (López & Strube, 2020), or learn label embeddings without explicit hierarchies (Chatterjee et al., 2021). Affective modeling has likewise used label-level geometric constraints to capture emotion dependencies (Yin et al., 2026).

These methods typically introduce task-specific label structures, whereas VMH-P derives structure directly from the class geometry of a trained host model. It converts this host-induced organization into structured prediction without modifying the underlying representation model.

2.3 Post-hoc Prediction Adaptation

Post-hoc adaptation decouples representation learning from decision optimization, for example by retraining only classifiers (Kang et al., 2020) or adjusting logits with frozen feature extractors (Menon et al., 2021). Similar ideas appear in vision–language adaptation: CLIP-Adapter (Gao et al., 2024) uses lightweight residual feature adaptation, whereas Tip-Adapter (Zhang et al., 2022) exploits cached features for efficient classification. These studies show that substantial predictive information remains exploitable after representation learning.

However, existing post-hoc methods primarily recalibrate logits, adapt features, or relearn representation-to-label mappings, without explicitly organizing relations among output categories. VMH-P extends this setting from prediction adaptation to structured prediction by inducing hierarchical and relational class organization from frozen multimodal representations and using the resulting evidence to refine the original host logits.

3 Method

This work introduces VMH-P, a plug-and-play structured prediction framework that derives complementary vertical and horizontal structures from the learned affective geometry of fully trained multimodal sentiment models, as shown in Fig. 1. VMH-P operates on representations extracted from a fixed host model, introducing structured prediction reasoning without modifying the original multimodal representation space.

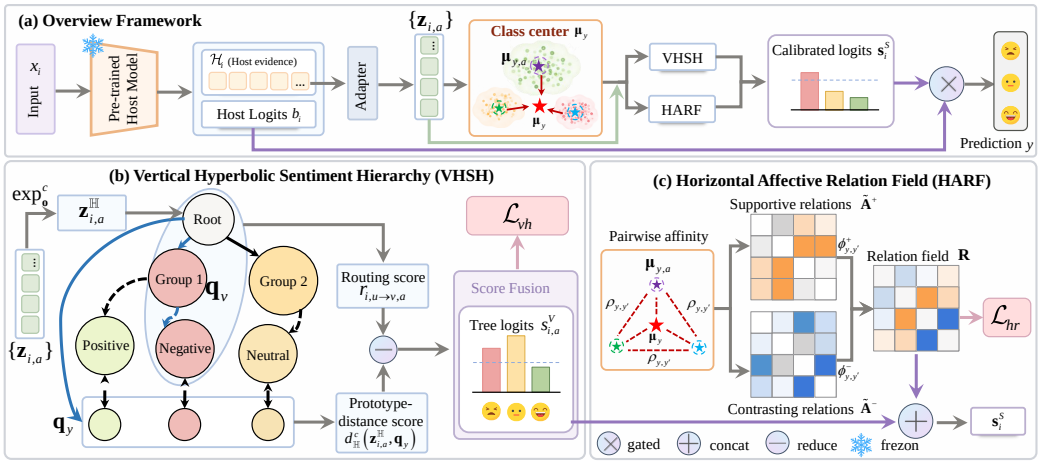


Figure 1: Overview of VMH-P. (a) Frozen host representations are aligned into a shared evidence space. (b) VHSH performs coarse-to-fine hyperbolic reasoning through routing and leaf discrimination. (c) HARF propagates evidence through refined supportive and contrasting relations.

3.1 POST-HOC PLUG-AND-PLAY PROTOCOL

Let $\mathcal{B}$ denote a trained multimodal sentiment model with fixed parameters $\theta^*_{\mathcal{B}}$. Let $\mathcal{A}$ denote the host evidence-source set, where $a \in \mathcal{A}$ indexes a textual, visual, or fused representation. Given an input x_i, the host model produces its original prediction logits $\mathbf{b}_i \in \mathbb{R}^C$ and a collection of intermediate evidence representations $\mathcal{H}_i$

$$(\mathbf{b}_i, \mathcal{H}_i) = \mathcal{B}(x_i; \theta^*_{\mathcal{B}}), \qquad \mathcal{H}_i = \{\mathbf{h}_{i,a} \mid a \in \mathcal{A}\}, \tag{1}$$

where C is the number of sentiment classes and $\mathbf{h}_{i,a}$ denotes the representation of the i-th sample from source a. A single forward pass over $\mathcal{D}_{\text{tr}}$ caches $\{(\mathbf{b}_i, \mathcal{H}_i, y_i)\}_{i \in \mathcal{D}_{\text{tr}}}$ for subsequent structure induction and VMH-P optimization, where y_i is the ground-truth label. Because host representations may differ in dimension and distribution, each source is mapped into a shared d-dimensional evidence space

$$\mathbf{z}_{i,a} = \mathrm{LN}\left(\mathbf{W}_a \mathbf{h}_{i,a} + \boldsymbol{\delta}_a\right), \qquad \mathbf{z}_{i,a} \in \mathbb{R}^d, \tag{2}$$

where $\mathbf{W}_a$ and $\boldsymbol{\delta}_a$ are source-specific projection parameters, and $\mathrm{LN}(\cdot)$ reduces distribution discrepancies across evidence sources. To establish consistent cross-source class geometry, source-specific centers are computed and aggregated into a shared class center

$$\boldsymbol{\mu}_{y,a} = \frac{\sum_{i \in \mathcal{D}_y} \mathbf{z}_{i,a}}{\left\| \sum_{i \in \mathcal{D}_y} \mathbf{z}_{i,a} \right\|_2 + \epsilon}, \qquad \boldsymbol{\mu}_y = \frac{\sum_{a \in \mathcal{A}} \boldsymbol{\mu}_{y,a}}{\left\| \sum_{a \in \mathcal{A}} \boldsymbol{\mu}_{y,a} \right\|_2 + \epsilon}, \tag{3}$$

where $\mathcal{D}_y \subset \mathcal{D}_{tr}$ denotes the subset of training samples belonging to class y, $\epsilon > 0$ is a small constant for numerical stability.

The evidence adapters are calibrated to preserve source-wise discriminability while aligning class geometry, using a shared auxiliary classifier $\mathbf{W}_e \in \mathbb{R}^{C \times d}$ only during calibration

$$\mathcal{L}_{adp} = \frac{1}{|\mathcal{D}_{tr}||\mathcal{A}|} \sum_{i \in \mathcal{D}_{tr}} \sum_{a \in \mathcal{A}} \mathrm{CE}\left(\mathbf{W}_e \mathbf{z}_{i,a}, y_i\right) + \frac{\lambda_{adp}}{C|\mathcal{A}|} \sum_{y=1}^{C} \sum_{a \in \mathcal{A}} \|\boldsymbol{\mu}_{y,a} - \boldsymbol{\mu}_y\|_2^2, \tag{4}$$

where $\mathrm{CE}(\cdot,\cdot)$ denotes cross-entropy and λ_{adp} balances source discrimination and class-geometry alignment. The two terms preserve source-wise discriminability and reduce cross-source geometric discrepancies, respectively. After calibration, the adapters are fixed, $\mathbf{W}_e$ is discarded, and the class centers are recomputed for subsequent structured reasoning.

3.2 VERTICAL HYPERBOLIC SENTIMENT HIERARCHY

The vertical axis of VMH-P captures coarse-to-fine class relations by inducing a task-adaptive hierarchy directly from the shared class geometry, without requiring hierarchical annotations.

The shared class centers $\{\boldsymbol{\mu}_y\}_{y=1}^C$ are partitioned into K latent affective groups

$$\{g_y\}_{y=1}^C = \mathrm{KMeans}\left(\{\boldsymbol{\mu}_y\}_{y=1}^C, K\right), \qquad g_y \in \{1, \ldots, K\}, \tag{5}$$

where $\mathrm{KMeans}(\cdot, K)$ denotes K-means clustering over the shared class centers, and g_y is the latent-group assignment of class y. These assignments define a three-level sentiment tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are the node and edge sets. It contains a global root, K latent affective regions, and C class leaves, forming a coarse-to-fine organization. Each internal node connects to the root, while class y is attached to the node indexed by g_y, allowing classes with similar host-induced representations to share a parent. The topology is induced once from the training-set geometry and remains fixed during structured reasoning.

To model the induced hierarchy, reasoning is performed in a Lorentz hyperbolic space

$$\mathbb{H}^d_c = \left\{ \mathbf{u} \in \mathbb{R}^{d+1} : \langle \mathbf{u}, \mathbf{u} \rangle_{\mathcal{L}} = -\frac{1}{c},\ u_0 > 0 \right\}, \qquad \mathbf{z}^H_{i,a} = \exp^c_o([0, \mathbf{z}_{i,a}]), \tag{6}$$

where $c > 0$ denotes the curvature magnitude, $[0, \mathbf{z}_{i,a}]$ denotes the tangent vector obtained by prepending a zero coordinate, $\exp^c_{\mathbf{o}}$ denotes the exponential map at the hyperbolic origin, and $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{L}} = -u_0 v_0 + \sum_{j=1}^d u_j v_j$ is the Lorentz inner product. The geodesic distance is $d^c_{\mathbb{H}}(\mathbf{u}, \mathbf{v}) = c^{-1/2} \operatorname{arcosh}(-c \langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{L}})$.

Each non-root node $v \in \mathcal{V} \setminus \{\text{root}\}$ is parameterized in the tangent space and mapped to $\mathbb{H}^d_c$ via the exponential map, keeping $\mathbf{q}_v$ on the Lorentz manifold. The root serves as a fixed structural anchor at the hyperbolic origin, while internal and leaf nodes are initialized from the corresponding class geometry

$$\mathbf{q}_k^{(0)} = \exp_{\mathbf{o}}^c\left(\left[0, \frac{1}{|\mathcal{Y}_k|}\sum_{y\in\mathcal{Y}_k}\boldsymbol{\mu}_y\right]\right), \qquad \mathbf{q}_y^{(0)} = \exp_{\mathbf{o}}^c\left([0, \boldsymbol{\mu}_y]\right), \tag{7}$$

where $\mathcal{Y}_k = \{y \mid g_y = k\}$ contains the classes assigned to group k. Each leaf $\mathbf{q}_y$ serves as both the terminal routing node and class prototype, avoiding separate prototype parameters.

Given $\mathbf{z}^{\mathbb{H}}_{i,a}$, the transition probability from a parent node u to a child $v \in \mathrm{Ch}(u)$ is

$$r_{i,u\to v,a} = \frac{\exp\left(-d^c_{\mathbb{H}}(\mathbf{z}^{\mathbb{H}}_{i,a}, \mathbf{q}_v)/\tau_r\right)}{\sum_{v'\in\mathrm{Ch}(u)}\exp\left(-d^c_{\mathbb{H}}(\mathbf{z}^{\mathbb{H}}_{i,a}, \mathbf{q}_{v'})/\tau_r\right)}, \tag{8}$$

where $\mathrm{Ch}(u)$ denotes the set of child nodes of u, and τ_r is the routing temperature. For class y, let $\mathcal{P}(y)$ denote the path from the root to its leaf.

Hierarchical routing may remain ambiguous among neighboring leaves, whereas leaf distance alone ignores intermediate structure. The vertical evidence score therefore combines path-level routing consistency with leaf-level discrimination

$$s^V_{i,y,a} = \sum_{(u\to v)\in\mathcal{P}(y)} \log r_{i,u\to v,a} - \beta d^c_{\mathbb{H}}\left(\mathbf{z}^{\mathbb{H}}_{i,a}, \mathbf{q}_y\right), \tag{9}$$

where the two terms capture path consistency and leaf-level discrimination, respectively, and β controls their relative contribution.

The induced hierarchy is preserved by constraining the relative geometry of leaf embeddings

$$\mathcal{L}_{vh} = 1 - \mathrm{Corr}_{y<y'}\left(D_{\mathcal{T}}(y, y'), d^c_{\mathbb{H}}(\mathbf{q}_y, \mathbf{q}_{y'})\right), \tag{10}$$

where $\mathrm{Corr}(\cdot,\cdot)$ is the Pearson correlation over class pairs and $D_{\mathcal{T}}(y, y')$ is their tree distance. Minimizing $\mathcal{L}_{\text{vh}}$ encourages hyperbolic distances to preserve the induced hierarchical geometry.

3.3 Horizontal Affective Relation Field

Since the vertical hierarchy does not model interactions at the same decision level, a horizontal relation field is introduced to capture supportive and contrasting relations among class leaves. Horizontal relations are initialized from the shared class centers $\boldsymbol{\mu}_y$. Unlike the vertical structure based on their spatial organization, horizontal interaction uses each class direction relative to the global centroid

$$\widehat{\boldsymbol{\mu}}_y = \frac{\boldsymbol{\mu}_y - \boldsymbol{\mu}_0}{\|\boldsymbol{\mu}_y - \boldsymbol{\mu}_0\|_2 + \epsilon}, \qquad \boldsymbol{\mu}_0 = \frac{1}{C}\sum_{y=1}^{C}\boldsymbol{\mu}_y, \tag{11}$$

where $\boldsymbol{\mu}_0$ is the global centroid and $\widehat{\boldsymbol{\mu}}_y$ is the normalized direction of class y. The pairwise angular affinity is then defined as

$$\rho_{y,y'} = \widehat{\boldsymbol{\mu}}_y^{\top}\widehat{\boldsymbol{\mu}}_{y'}, \qquad y \neq y', \tag{12}$$

where positive and negative values indicate aligned and contrasting class directions, respectively.

The signed affinities initialize the supportive and contrasting relation components as

$$A^{+(0)}_{y,y'} = \frac{[\rho_{y,y'}]_+\,\mathbb{I}[y\neq y']}{\sum_{k\neq y}[\rho_{y,k}]_+ + \epsilon}, \qquad A^{-(0)}_{y,y'} = \frac{[-\rho_{y,y'}]_+\,\mathbb{I}[y\neq y']}{\sum_{k\neq y}[-\rho_{y,k}]_+ + \epsilon}, \tag{13}$$

where $A^{+(0)}_{y,y'}$ and $A^{-(0)}_{y,y'}$ denote the initial supportive and contrasting relation strengths from class y to class y', respectively. Here, $[x]_+ = \max(x, 0)$, $\mathbb{I}[\cdot]$ denotes the indicator function used to remove self-relations, and k indexes the sentiment classes involved in row-wise normalization.

Fixing the geometry-derived relations would prevent adaptation to the structured prediction objective, whereas unconstrained relation learning could discard the class geometry inherited from the

host. The initial relations are therefore retained as a structural prior, with their topology and signs preserved while their strengths are adaptively refined during VMH-P optimization

$$\widetilde{A}^{+}_{y,y'} = \frac{A^{+(0)}_{y,y'} \exp\left(\phi^{+}_{y,y'}\right)}{\sum_{k\neq y} A^{+(0)}_{y,k} \exp\left(\phi^{+}_{y,k}\right) + \epsilon}, \qquad \widetilde{A}^{-}_{y,y'} = \frac{A^{-(0)}_{y,y'} \exp\left(\phi^{-}_{y,y'}\right)}{\sum_{k\neq y} A^{-(0)}_{y,k} \exp\left(\phi^{-}_{y,k}\right) + \epsilon}, \tag{14}$$

where $\phi^{+}_{y,y'}$ and $\phi^{-}_{y,y'}$ are trainable relation-refinement parameters initialized to zero. Collecting the pairwise relations yields $\mathbf{A}^{+(0)}, \mathbf{A}^{-(0)}, \widetilde{\mathbf{A}}^{+}, \widetilde{\mathbf{A}}^{-} \in \mathbb{R}^{C\times C}$. The refined signed relation operator is

$$\mathbf{R} = \widetilde{\mathbf{A}}^{+} - \widetilde{\mathbf{A}}^{-}. \tag{15}$$

To prevent excessive deviation from the geometry-derived initialization, horizontal relation consistency is encouraged by

$$\mathcal{L}_{hr} = \left\| \widetilde{\mathbf{A}}^{+} - \mathbf{A}^{+(0)} \right\|_F^2 + \left\| \widetilde{\mathbf{A}}^{-} - \mathbf{A}^{-(0)} \right\|_F^2, \tag{16}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This formulation allows the horizontal relations to be prediction-adaptive while retaining the inter-class geometry supplied by the frozen host.

Class-wise evidence vectors are treated as row vectors in the following relation propagation. Let $\mathbf{s}^{V}_{i,a} = \{s^{V}_{i,y,a}\}_{y=1}^{C}$ denote the vertical evidence scores of source a. The vertical scores are first normalized into class-wise evidence, which is then propagated across sentiment classes according to the refined horizontal relations:

$$\boldsymbol{\pi}^{V}_{i,a} = \Pi\left(\mathbf{s}^{V}_{i,a}/\tau_e\right), \qquad \mathbf{s}^{R}_{i,a} = \mathbf{s}^{V}_{i,a} + \eta \boldsymbol{\pi}^{V}_{i,a}\mathbf{R}, \tag{17}$$

where $\Pi(\cdot)$ denotes the softmax operation, τ_e is the evidence temperature, and η controls the strength of horizontal relation propagation. Here, $\boldsymbol{\pi}^{V}_{i,a}$ represents the normalized class-wise evidence produced by vertical reasoning, while $\boldsymbol{\pi}^{V}_{i,a}\mathbf{R}$ aggregates the influence transferred from all source classes through the horizontal relation field. Specifically, the propagated evidence received by class y' is $\sum_{y=1}^{C} \pi^{V}_{i,y,a} R_{y,y'}$, where $R_{y,y'}$ denotes the refined relation from class y to class y'. Consequently, evidence associated with supportive relations reinforces geometrically aligned classes, whereas evidence transmitted through contrasting relations suppresses oppositely oriented classes. The residual form preserves the original vertical evidence and introduces horizontal interaction only as an additional correction.

Since different evidence sources exhibit sample-dependent reliability, the relation-enhanced predictions are further aggregated according to their uncertainty. Let the confidence of source a be $\kappa_{i,a} = 1 - \mathrm{Ent}\left(\Pi(\mathbf{s}^{R}_{i,a}/\tau_e)\right) / \log C$, where $\mathrm{Ent}(\cdot)$ denotes entropy. The final structured logits are obtained by

$$\omega_{i,a} = \frac{\exp(\kappa_{i,a}/\tau_s)}{\sum_{a'\in\mathcal{A}} \exp(\kappa_{i,a'}/\tau_s)}, \qquad \mathbf{s}^{S}_{i} = \sum_{a\in\mathcal{A}} \omega_{i,a}\mathbf{s}^{R}_{i,a}, \tag{18}$$

where τ_s is the source-selection temperature. Thus, evidence sources with lower predictive uncertainty contribute more strongly to the final structured prediction.

Since the reliability of structured reasoning may vary across samples, its contribution is controlled by the relative confidence between the structured and host predictions. Structured evidence is then introduced through a confidence-gated residual correction

$$\alpha_i = \sigma\left(\frac{\mathrm{Ent}(\Pi(\mathbf{b}_i)) - \mathrm{Ent}\left(\Pi(\mathbf{s}^{S}_{i})\right)}{\tau_g \log C}\right), \qquad \mathbf{o}_i = \mathbf{b}_i + \gamma\alpha_i\mathbf{s}^{S}_{i}, \tag{19}$$

where $\sigma(\cdot)$ denotes the sigmoid function, τ_g controls the gate sensitivity, γ is a learnable residual scale initialized to zero, and $\mathbf{o}_i$ denotes the final prediction logits. A more confident structured prediction yields a larger α_i, while γ controls the correction magnitude and its zero initialization preserves the original prediction.

The structured branch and final prediction are jointly optimized by

$$\mathcal{L}_{VMH} = \frac{1}{|\mathcal{D}_{tr}|} \sum_{i\in\mathcal{D}_{tr}} \left[\mathrm{CE}(\mathbf{o}_i, y_i) + \lambda_{str}\mathrm{CE}(\mathbf{s}^{S}_{i}, y_i)\right] + \lambda_{vh}\mathcal{L}_{vh} + \lambda_{hr}\mathcal{L}_{hr}, \tag{20}$$

where λ_{str}, λ_{vh}, and λ_{hr} control structured prediction supervision, vertical hierarchy consistency, and horizontal relation consistency, respectively.

4 EXPERIMENTS

4.1 SETUP

Datasets. VMH-P is evaluated on MVSA-Single and MVSA-Multiple (Niu et al., 2016), containing 4,511 and 17,024 image–text pairs with three polarity labels, and TumEmo (Yang et al., 2021), containing 195,265 pairs with seven emotion categories. All datasets use an 80%/10%/10% train/validation/test split, with Accuracy (ACC) and Weighted-F1 (F1) as evaluation metrics.

Host Models. To evaluate backbone-agnostic plug-and-play capability, CLMLF (Li et al., 2022), D2R (Chen et al., 2024), and SPP-SCL (Wu & Li, 2026) are adopted as three heterogeneous host models. CLMLF combines multi-layer fusion with contrastive learning, D2R performs sample-adaptive multimodal interaction through dynamic routing, and SPP-SCL balances intra- and inter-modal sentiment representations. Each host is fully trained beforehand and kept frozen during VMH-P optimization. The available multimodal representations exposed by each host, together with its original logits, are cached and used as inputs to VMH-P without modifying the host architecture.

Implementation Details. VMH-P is optimized only on cached host representations. Set $d = 64$, $c = 0.25$, $\tau_r = 0.07$, and $\beta = 0.30$, with $K = 2$ for MVSA-Single and MVSA-Multiple and $K = 3$ for TumEmo. Other hyperparameters are $\lambda_{adp} = 0.1$, $\lambda_{vh} = 0.10$, $\lambda_{hr} = 0.01$, $\lambda_{str} = 0.10$, $\eta = 0.50$, $\tau_e = 0.10$, $\tau_s = 0.10$, and $\tau_g = 0.10$. AdamW uses a weight decay of 10^{-4}. The adapter and structured-head learning rates are 1×10^{-4} and 3×10^{-4}, respectively. Training runs for up to 10 epochs with batch size 64 and early-stopping patience 2. Experiments use one NVIDIA RTX 5090 GPU with three seeds. Following the evaluation protocol of the compared methods, the main results report the best run, with mean$\pm$std results provided in Appendix A.3. All variants use identical cached representations, data splits, and optimization protocols for fair comparison.

4.2 MAIN RESULTS

Table 1 shows that VMH-P improves all host–dataset combinations, with average gains of 3.84% and 4.12% in ACC and F1, respectively. The improvements also exhibit clear host-dependent patterns. For CLMLF, the gain is particularly pronounced on TumEmo. Its multi-layer fusion and contrastive learning encourage cross-modal representation alignment, while the resulting fine-grained affective geometry may remain only partially exploited by the original flat classifier. VMH-P explicitly organizes such latent class relations through vertical hierarchy and horizontal interaction, providing complementary prediction refinement.

SPP-SCL, in contrast, shows especially strong gains on MVSA-Single and MVSA-Multiple. Its supervised contrastive objective encourages discriminative intra- and inter-modal sentiment representations, which may form more structured class geometry for polarity-oriented categories. VMH-P further converts this geometry into hierarchical and relational evidence that is not explicitly utilized by the original prediction head. Together with the consistent gains on D2R, these results indicate that VMH-P exploits affective geometry beyond flat classification heads, supporting its plug-and-play capability across heterogeneous hosts.

4.3 ABLATION STUDIES

Comparison with Generic Post-hoc Heads. To test whether the gains merely arise from added prediction capacity, three alternatives are evaluated on the same cached representations. Linear Residual applies linear logit correction, MLP Residual uses a two-layer predictor, and Euclidean Prototype classifies by distances to prototypes from cached training features. As shown in Table 2, Linear and MLP Residual yield marginal gains, whereas Euclidean Prototype performs better without trainable parameters, highlighting the value of class geometry over added prediction capacity.

Effect of Vertical and Horizontal Structures. Table 3 evaluates five variants. *Euclidean VMH-P* replaces the Lorentz hierarchy with a Euclidean counterpart, while *w/o Horizontal* and *w/o Vertical* remove horizontal propagation and hierarchical routing, respectively. *Random Tree* and *Random Relation* randomize the two structures. Randomizing either consistently degrades performance, indicating that the recovered hierarchy and relation field capture meaningful prediction cues rather than arbitrary bias. Euclidean VMH-P also performs worse across all host–dataset combinations,

Table 1: Main results. Underlines mark best competitors; parentheses show gains over frozen hosts.

Model	MVSA-Single		MVSA-Multiple		TumEmo	
	ACC	F1	ACC	F1	ACC	F1
M2CL (Yang et al., 2023)	75.50	74.20	73.20	70.50	–	–
SHSL (Han et al., 2024)	75.72	75.20	71.20	70.39	68.57	68.60
CTMWA (Zhang et al., 2024)	75.91	75.74	74.02	73.84	68.57	68.60
DRF (Wu et al., 2024)	74.50	74.40	71.00	68.20	69.60	69.60
MAMSA (Wang et al., 2025b)	77.61	76.84	–	–	68.06	67.92
HIFCL (Wang et al., 2025a)	77.01	76.18	–	–	68.44	68.27
CoDe (Wu et al., 2026)	76.98	76.47	72.74	70.71	73.02	73.02
TiCAL (Yin et al., 2026)	73.17	73.09	67.58	66.04	73.87	73.86
SentiFact (You et al., 2027)	76.97	75.58	69.81	69.23	70.03	69.98
CLMLF (Li et al., 2022)	75.33	73.46	72.00	69.83	68.10	68.00
+ VMH-P	**78.22** (+2.89)	**77.39** (+3.93)	**75.53** (+3.53)	**73.57** (+3.74)	**73.35** (+5.25)	**73.32** (+5.32)
D2R (Chen et al., 2024)	76.67	75.59	71.59	70.85	<u>74.64</u>	<u>74.60</u>
+ VMH-P	**78.00** (+1.33)	**77.94** (+2.35)	**76.12** (+4.53)	**74.26** (+3.41)	**77.39** (+2.75)	**77.41** (+2.81)
SPP-SCL (Wu & Li, 2026)	<u>81.33</u>	<u>80.15</u>	<u>78.71</u>	<u>77.53</u>	72.15	72.13
+ VMH-P	**89.56** (+8.23)	**89.04** (+8.89)	**83.35** (+4.64)	**82.66** (+5.13)	**73.59** (+1.44)	**73.67** (+1.54)

Table 2: Comparison with generic post-hoc heads. Performance gains are averaged over all host–dataset combinations, while parameter counts indicate the corresponding range.

Method	Params (M)	ΔACC	ΔF1
Linear Residual	0.0004–0.0161	+0.05	+0.05
MLP Residual	0.0341–0.5965	+0.59	+0.74
Euclidean Prototype	0	+0.90	+1.39
VMH-P	0.0128–0.4746	**+3.84**	**+4.12**

supporting hyperbolic geometry for coarse-to-fine hierarchy modeling. Finally, removing either vertical or horizontal reasoning degrades performance, confirming their complementary roles.

Table 3: Ablation of vertical and horizontal structures and comparison to the Euclidean counterpart.

Host	Dataset	Full VMH-P	Euclidean VMH-P	w/o Horizontal	w/o Vertical	Random Tree	Random Relation
CLMLF	MVSA-S	**78.22/77.39**	77.33/76.55	76.00/74.80	75.56/74.67	75.78/74.46	76.00/74.85
	MVSA-M	**75.53/73.57**	74.71/72.45	73.00/70.05	72.47/70.35	72.65/69.90	72.59/69.86
	TumEmo	**73.35/73.32**	71.90/71.72	70.80/70.62	71.19/70.91	70.90/70.76	70.71/70.59
D2R	MVSA-S	**78.00/77.94**	77.11/76.90	76.56/76.56	76.89/76.45	76.33/75.91	76.11/76.00
	MVSA-M	**76.12/74.26**	74.35/74.11	72.94/71.04	72.18/70.86	73.71/71.14	72.76/71.09
	TumEmo	**77.39/77.41**	76.45/76.48	74.97/74.99	75.10/75.12	74.91/74.93	74.96/74.99
SPP-SCL	MVSA-S	**89.56/89.04**	87.33/86.83	86.89/85.37	86.89/85.75	87.11/86.00	86.89/85.37
	MVSA-M	**83.35/82.66**	81.65/81.11	79.00/78.00	81.88/81.57	79.24/77.69	78.94/77.84
	TumEmo	**73.59/73.67**	72.26/72.35	72.38/72.44	72.51/72.57	72.39/72.47	72.38/72.44

4.4 Analysis of Induced Class Structure

Figs. 2 and 3 visualize the vertical hierarchies and horizontal relation fields induced on TumEmo. The similar tree topologies suggest shared coarse affective organization across hosts, while different routing probabilities reflect host-specific evidence geometry. Notably, Calm is grouped with Happy and Love despite semantic differences, reflecting host-induced organization rather than a predefined taxonomy. These results show that VMH-P maintains task-level structural consistency while adapting coarse-to-fine reasoning to each learned affective space.

Figure 3 shows that VMH-P induces supportive and contrasting interactions whose strengths vary across hosts while retaining clear positive and negative patterns. This host-dependent structure indicates non-uniform evidence propagation and reflects the adaptive refinement of relations initialized from class directions. Together, the visualizations show complementary roles of the two structures.

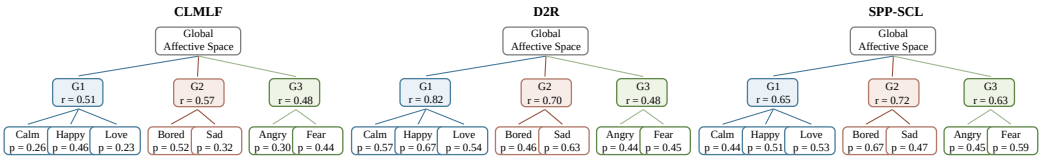


Figure 2: Induced vertical hierarchies on TumEmo for three hosts. r is the region-conditioned mean root-to-region probability, and p the class-conditioned mean root-to-leaf probability.

Vertical reasoning organizes classes across levels of granularity, whereas horizontal reasoning models interactions at the same decision level. VMH-P thus converts learned affective geometry into prediction-oriented structure without imposing a predefined sentiment organization.

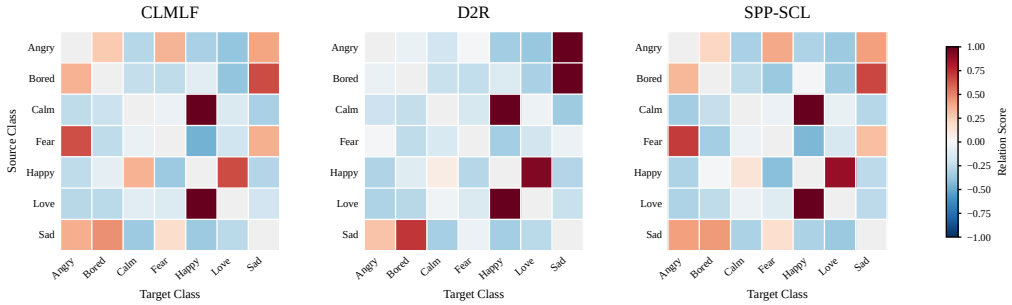


Figure 3: Induced horizontal relation fields on TumEmo across three host models. Positive and negative values denote supportive and contrasting relations, respectively.

4.5 Efficiency and Plug-and-Play Analysis

VMH-P operates only on cached representations without updating the frozen backbone, avoiding repeated host optimization. As shown in Table 4, it introduces at most 0.4746M trainable parameters, 5.43×10^{-4} GFLOPs, and 0.0206 ms latency per sample. The variation across hosts mainly stems from differences in exposed evidence representations and label spaces, while the vertical and horizontal class reasoning remains lightweight. These results confirm its low overhead and practical plug-and-play capability.

Table 4: Efficiency of VMH-P across three hosts. Values are ranges over three datasets.

Host	Params (M)	GFLOPs	Extra Time (ms/sample)
D2R	0.1521–0.1530	1.60×10^{-4}–1.69×10^{-4}	0.0120–0.0178
CLMLF	0.1521–0.4746	1.63×10^{-4}–5.43×10^{-4}	0.0118–0.0206
SPP-SCL	0.0128–0.0139	2.35×10^{-5}–3.04×10^{-5}	0.0077–0.0171

5 Conclusion

This work presents VMH-P, a plug-and-play post-hoc structured prediction head for multimodal sentiment analysis. VMH-P preserves the learned representation space while disentangling learned affective geometry into complementary vertical and horizontal decision structures. Vertical hyperbolic reasoning supports coarse-to-fine routing and leaf discrimination, whereas horizontal relation reasoning propagates evidence through supportive and contrasting class interactions. The resulting structured evidence is introduced into the host prediction through confidence-gated residual correction. Experiments across three multimodal affective datasets and three heterogeneous hosts show consistent improvements across all host–dataset combinations, with average gains of 3.84% and 4.12% in ACC and F1. These results demonstrate the effectiveness of exploiting class organization at the prediction stage and highlight the potential of structured prediction as a lightweight extension to existing multimodal models.

AI USE STATEMENT

Generative AI tools were used only for language polishing to improve grammar, clarity, and conciseness. All AI-assisted revisions were reviewed by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

To support reproducibility, the main paper specifies the datasets, host models, evaluation metrics, training protocol, and key hyperparameters used in all experiments. The appendix provides additional implementation details, host-specific evidence sources, hyperparameter settings, ablation configurations, and analyses of the recovered class structures. All comparisons use identical data splits, cached representations, and optimization protocols where applicable. The anonymous source code is available through the link provided in the abstract.

REFERENCES

Jieyu An and Wan Mohd Nazmee Wan Zainon. Integrating color cues to improve multimodal sentiment analysis in social media. *Engineering Applications of Artificial Intelligence*, 126: 106874, 2023. ISSN 0952-1976. doi: 10.1016/j.engappai.2023.106874. URL https://www.sciencedirect.com/science/article/pii/S0952197623010588.

Soumya Chatterjee, Ayush Maheshwari, Ganesh Ramakrishnan, and Saketha Nath Jagarlapudi. Joint learning of hyperbolic label embeddings for hierarchical multi-label classification. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2829–2841, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.247. URL https://aclanthology.org/2021.eacl-main.247/.

Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. In *Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding*, MMPT '21, pp. 37–45, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385305. doi: 10.1145/3463945.3469058. URL https://doi.org/10.1145/3463945.3469058.

Boli Chen, Xin Huang, Lin Xiao, Zixin Cai, and Liping Jing. Hyperbolic interaction model for hierarchical multi-label classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7496–7503. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6247. URL https://doi.org/10.1609/aaai.v34i05.6247.

Yifan Chen, Kuntao Li, Weixing Mai, Qiaofeng Wu, Yun Xue, and Fenghuan Li. D2R: dual-branch dynamic routing network for multimodal sentiment detection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 3536–3547. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.207. URL https://doi.org/10.18653/v1/2024.emnlp-main.207.

Alan S. Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017. doi: 10.1073/pnas.1702247114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1702247114.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.*, 132(2):581–595, 2024. doi: 10.1007/S11263-023-01891-X. URL https://doi.org/10.1007/s11263-023-01891-x.

Xue Han, Honlin Cheng, Jike Ding, and Suqin Yan. Semisupervised hierarchical subspace learning model for multimodal social media sentiment analysis. *IEEE Trans. Consumer Electron.*, 70(1):

3446–3454, 2024. doi: 10.1109/TCE.2024.3350696. URL https://doi.org/10.1109/TCE.2024.3350696.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. CLMLF: A contrastive learning and multilayer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2282–2294, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/V1/2022.FINDINGS-NAACL.175. URL https://doi.org/10.18653/V1/2022.FINDINGS-NAACL.175.

Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 7124–7135, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.622/.

Federico López and Michael Strube. A fully hyperbolic neural model for hierarchical multi-class classification. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 460–475, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.42. URL https://aclanthology.org/2020.findings-emnlp.42/.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6338–6347, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html.

Teng Niu, Shuhui Zhu, Liangliang Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In Qing Tian, Nicu Sebe, Guo-Jun Qi, Benoit Huet, Richang Hong, and Xian-Sheng Hua (eds.), *MultiMedia Modeling (MMM 2016)*, volume 9517 of *Lecture Notes in Computer Science*, pp. 15–27. Springer, Cham, 2016. doi: 10.1007/978-3-319-27674-8_2. URL https://doi.org/10.1007/978-3-319-27674-8_2.

Alfonso Semeraro, Salvatore Vilella, and Giancarlo Ruffo. Pyplutchik: visualising and comparing emotion-annotated corpora. *CoRR*, abs/2105.04295, 2021. URL https://arxiv.org/abs/2105.04295.

Hongbin Wang, Qifei Du, and Yan Xiang. Image–text sentiment analysis based on hierarchical interaction fusion and contrast learning enhanced. *Engineering Applications of Artificial Intelligence*, 146:110262, 2025a. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2025.110262. URL https://www.sciencedirect.com/science/article/pii/S0952197625002623.

Hongbin Wang, Chun Ren, and Zhengtao Yu. Multimodal sentiment analysis based on multiple attention. *Eng. Appl. Artif. Intell.*, 140:109731, 2025b. doi: 10.1016/J.ENGAPPAI.2024.109731. URL https://doi.org/10.1016/j.engappai.2024.109731.

Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3664647.3681403. URL https://doi.org/10.1145/3664647.3681403.

Daiqing Wu, Dongbao Yang, Yu Zhou, and Can Ma. Robust multimodal sentiment analysis of image-text pairs by distribution-based feature recovery and fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 5780–5789, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3680653. URL https://doi.org/10.1145/3664647.3680653.

Daiqing Wu, Dongbao Yang, Huawen Shen, Can Ma, and Yu Zhou. Resolving sentiment discrepancy for multimodal sentiment detection via semantics completion and decomposition. *Pattern Recognit.*, 172:112719, 2026. doi: 10.1016/J.PATCOG.2025.112719. URL https://doi.org/10.1016/j.patcog.2025.112719.

Jiesheng Wu and Shengrong Li. SPP-SCL: semi-push-pull supervised contrastive learning for image-text sentiment analysis and beyond. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 2173–2181. AAAI Press, 2026. doi: 10.1609/AAAI.V40I3.37200. URL https://doi.org/10.1609/aaai.v40i3.37200.

Qinfu Xu, Yiwei Wei, Chunlei Wu, Leiquan Wang, Shaozu Yuan, Jie Wu, Jing Lu, and Hengyang Zhou. Towards multimodal sentiment analysis via hierarchical correlation modeling with semantic distribution constraints. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 21788–21796. AAAI Press, 2025. doi: 10.1609/AAAI.V39I20.35484. URL https://doi.org/10.1609/aaai.v39i20.35484.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Trans. Multim.*, 23:4014–4026, 2021. doi: 10.1109/TMM.2020.3035277. URL https://doi.org/10.1109/TMM.2020.3035277.

Xiaocui Yang, Shi Feng, Daling Wang, Pengfei Hong, and Soujanya Poria. Multiple contrastive learning for multimodal sentiment analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096777.

Wen Yin, Siyu Zhan, Cencen Liu, Xin Hu, Guiduo Duan, Xiurui Xie, Yuan-Fang Li, and Tao He. Tical: Typicality-based consistency-aware learning for multimodal emotion recognition. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 17948–17956. AAAI Press, 2026. doi: 10.1609/AAAI.V40I21.38854. URL https://doi.org/10.1609/aaai.v40i21.38854.

Jiali You, Haoran Li, Jiawen Deng, Wei Li, Yuanyuan He, and Fuji Ren. Salience-guided counterfactual framework for multimodal sentiment analysis. *Information Fusion*, 137:104587, 2027. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2026.104587. URL https://www.sciencedirect.com/science/article/pii/S1566253526004653.

Jianfei Yu and Jing Jiang. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pp. 5408–5414, Macao, China, August 2019. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2019/751. URL https://doi.org/10.24963/ijcai.2019/751.

Jianfei Yu, Kai Chen, and Rui Xia. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):1966–1978, 2023. doi: 10.1109/TAFFC.2022.3171091.

Baozheng Zhang, Ziqi Yuan, Hua Xu, and Kai Gao. Crossmodal translation based meta weight adaption for robust image-text sentiment analysis. *IEEE Trans. Multim.*, 26:9949–9961, 2024. doi: 10.1109/TMM.2024.3405662. URL https://doi.org/10.1109/TMM.2024.3405662.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of CLIP for few-shot classification. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV*, volume 13695 of *Lecture Notes in Computer Science*, pp. 493–510. Springer, 2022. doi: 10.1007/978-3-031-19833-5_29. URL https://doi.org/10.1007/978-3-031-19833-5_29.

A Additional Implementation Details

A.1 Host-Specific Evidence Extraction

VMH-P operates on intermediate representations already produced by a fully trained host model and does not modify its original representation-learning architecture. For each host, the available textual, visual, and multimodal representations are extracted during a single forward pass and cached together with the original prediction logits and ground-truth labels. These cached representations are subsequently used for evidence alignment, class-structure induction, and structured-head optimization.

The three host models expose different representation pathways. CLMLF performs multi-layer cross-modal fusion with contrastive learning, D2R dynamically routes multimodal interactions through dual branches, and SPP-SCL aligns intra- and inter-modal sentiment representations before fusion. VMH-P does not require these architectures to expose identical internal modules. Instead, each available evidence source is independently projected into the shared d-dimensional evidence space through a source-specific linear adapter followed by LayerNorm. This source-wise adaptation allows VMH-P to accommodate heterogeneous feature dimensions and distributions while maintaining the same structured prediction pipeline across hosts.

Importantly, all host parameters remain frozen throughout VMH-P training. No gradient is propagated into the original multimodal backbone, and removing VMH-P directly recovers the original host prediction without any parameter restoration or architectural modification.

A.2 Training and Optimization Details

VMH-P is optimized after the host model has been fully trained and frozen. The evidence adapters are first calibrated on cached training representations. During calibration, the shared class centers used as alignment targets are recomputed from the current adapted training representations at the beginning of each epoch and kept fixed within that epoch, while the source-specific class centers are computed from the current adapted representations during optimization and retain gradients with respect to the evidence adapters. After calibration, the adapters are frozen, and the class centers are recomputed once for subsequent structure induction and structured-head optimization. The induced tree topology and geometry-derived relation field are then constructed from these calibrated class centers, while no gradient is propagated into the host model. Table 5 summarizes the hyperparameters used in our experiments.

A.3 Multi-Seed Stability

To assess robustness to random initialization, VMH-P is trained with three seeds (42, 43, and 44) under identical frozen hosts, cached representations, data splits, and hyperparameters. As shown in Table 6, VMH-P maintains consistent gains with generally small standard deviations across host–dataset combinations. The slightly larger variation of SPP-SCL on MVSA-Multiple suggests greater

Table 5: Hyperparameter settings of VMH-P.

Category	Hyperparameter	Value
Evidence Adapter	Shared evidence dimension d	64
	Adapter alignment weight λ_{adp}	0.10
Vertical Structure	Number of latent groups K	2 (MVSA-S/M), 3 (TumEmo)
	Curvature magnitude c	0.25
	Routing temperature τ_r	0.07
	Leaf discrimination weight β	0.30
	Hierarchy consistency weight λ_{vh}	0.10
Horizontal Structure	Relation regularization weight λ_{hr}	0.01
	Relation propagation strength η	0.50
	Evidence temperature τ_e	0.10
	Source-selection temperature τ_s	0.10
Prediction	Structured loss weight λ_{str}	0.10
	Gating temperature τ_g	0.10
Optimization	Optimizer	AdamW
	Adapter learning rate	1×10^{-4}
	Structured-head learning rate	3×10^{-4}
	Weight decay	1×10^{-4}
	Batch size	64
	Maximum epochs	10
	Early-stopping patience	2

Table 6: Multi-seed stability of VMH-P over three random seeds (42, 43, and 44). Results are reported as mean $\pm$ standard deviation.

Host	Dataset	Frozen Host (ACC/F1)	VMH-P ACC	VMH-P F1
CLMLF	MVSA-Single	75.33/73.46	77.70 ± 0.71	76.99 ± 0.52
	MVSA-Multiple	72.00/69.83	75.00 ± 0.53	73.17 ± 0.44
	TumEmo	68.10/68.00	73.02 ± 0.28	72.91 ± 0.36
D2R	MVSA-Single	76.67/75.59	77.63 ± 0.46	77.32 ± 0.75
	MVSA-Multiple	71.59/70.85	75.84 ± 0.27	74.20 ± 0.08
	TumEmo	74.64/74.60	77.35 ± 0.05	77.37 ± 0.05
SPP-SCL	MVSA-Single	81.33/80.15	89.04 ± 0.46	88.47 ± 0.50
	MVSA-Multiple	78.71/77.53	82.51 ± 1.21	81.70 ± 1.35
	TumEmo	72.15/72.13	73.54 ± 0.08	73.61 ± 0.08

sensitivity of its induced geometry to initialization, while the overall improvements remain stable across runs.

A.4 Hyperparameter Sensitivity

We further analyze the sensitivity of VMH-P to four representative hyperparameters on TumEmo in terms of ACC: the number of latent groups K, curvature magnitude c, relation propagation strength η, and structured loss weight λ_{str}. Each parameter is varied independently while keeping the remaining settings fixed. As shown in Figure 4, performance remains relatively stable over a broad range of configurations across different host models. Notably, the default configuration is not always the individually optimal choice for every host, suggesting that the main results do not rely on host-specific hyperparameter tuning.

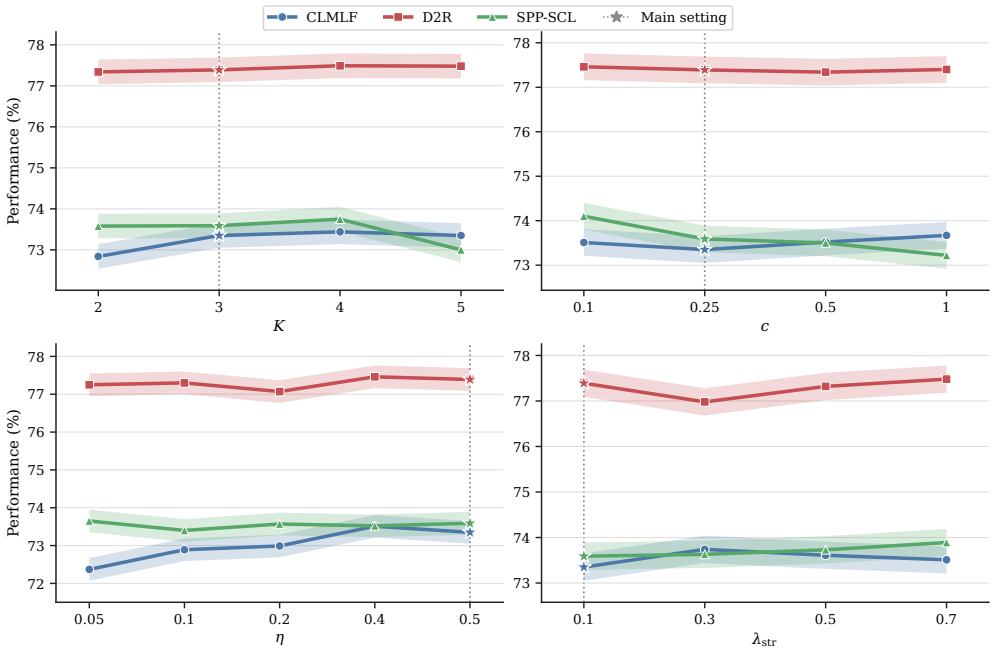


Figure 4: Hyperparameter sensitivity of VMH-P with respect to the number of latent groups K, curvature magnitude c, relation propagation strength η, and structured loss weight λ_{str}. Each hyperparameter is varied independently while keeping the others fixed.

B Additional Analysis of Recovered Structures

B.1 Vertical Hierarchy Visualization

Figure 5 presents the vertical hierarchies induced on MVSA-Single and MVSA-Multiple across the three hosts. For each host, the two datasets exhibit a consistent coarse grouping: CLMLF groups Negative with Neutral, D2R groups Positive with Neutral, while SPP-SCL groups Negative with Positive. These different tree topologies arise from the host-specific geometry of the learned affective space. Specifically, VMH-P constructs the hierarchy by clustering class centers derived from the frozen representations of each host. Since different host models learn different representation spaces and therefore different relative arrangements among sentiment categories, the nearest class centers in one host are not necessarily the nearest in another. Consequently, the induced hierarchy reflects how each host organizes sentiment evidence in its own representation space, rather than imposing a predefined semantic taxonomy. Meanwhile, the routing probabilities further vary across hosts and datasets, showing that VMH-P adapts not only the coarse class grouping but also the strength of coarse-to-fine evidence allocation. Together with the TumEmo results in the main paper, these visualizations demonstrate that VMH-P induces host-dependent class organization from learned affective geometry and converts it into adaptive hierarchical prediction.

B.2 Horizontal Relation Visualization

Figure 6 visualizes the horizontal relation fields induced on MVSA-Single and MVSA-Multiple across the three hosts. Positive and negative values denote supportive and contrasting relations, respectively. The differences across hosts arise from their host-specific affective geometries. Specifically, VMH-P initializes horizontal relations according to the relative directions of class centers with respect to the global centroid. Since different host models organize sentiment classes differently in their learned representation spaces, the pairwise angular affinities among classes also vary, leading to different supportive and contrasting relation patterns. These geometry-derived relations are further refined during structured prediction while remaining constrained by the original host-induced geometry. Consequently, the resulting relation fields capture host-dependent patterns of inter-class

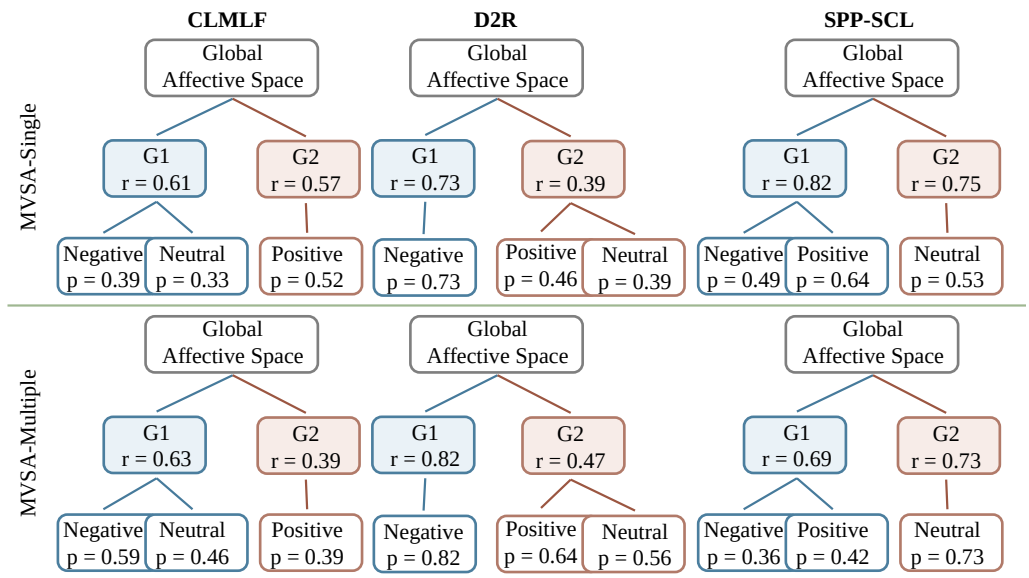


Figure 5: Induced vertical hierarchies on MVSA-Single and MVSA-Multiple across three hosts. r and p denote the mean root-to-region and root-to-leaf path routing probabilities, respectively.

evidence propagation rather than a fixed semantic relation graph shared across models. Together with the TumEmo example in the main paper, these results show that VMH-P adapts horizontal class interactions to the geometry learned by each host.

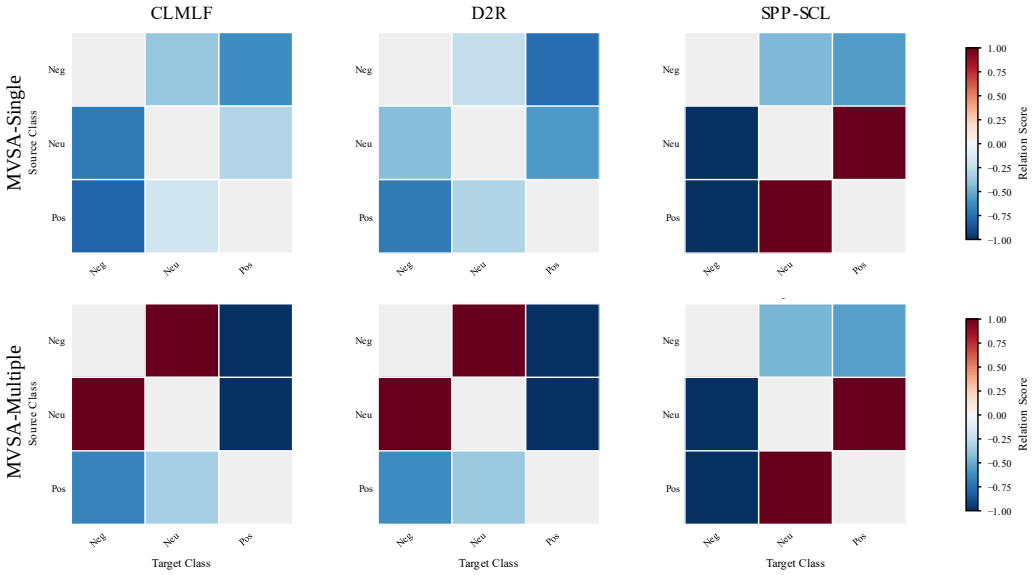


Figure 6: Induced horizontal relation fields on MVSA-Single and MVSA-Multiple across three host models. Positive and negative values denote supportive and contrasting relations, respectively.

C Case Study

Figure 7 presents two representative TumEmo examples to illustrate how the vertical and horizontal structures contribute to prediction refinement.

In Case (a), the frozen D2R host incorrectly predicts *Sad* with a confidence of 0.65, while assigning only 0.08 to the ground-truth class *Angry*. VMH-P reorganizes the evidence through coarse-to-fine hierarchical routing, shifting the prediction toward the affective region containing *Angry* and yielding a final confidence of 0.91. This case illustrates how vertical reasoning corrects an inter-region error.

Case (b) represents a finer intra-region ambiguity between *Happy* and *Love*. Although the frozen host favors *Happy* (0.52), vertical reasoning assigns nearly equal leaf evidence to *Happy* and *Love*. The positive horizontal relation from *Happy* to *Love* then transfers additional evidence toward *Love*, leading VMH-P to predict *Love* with a confidence of 0.53. This example shows that horizontal reasoning complements the vertical hierarchy by resolving ambiguity between closely related classes within the same affective region.

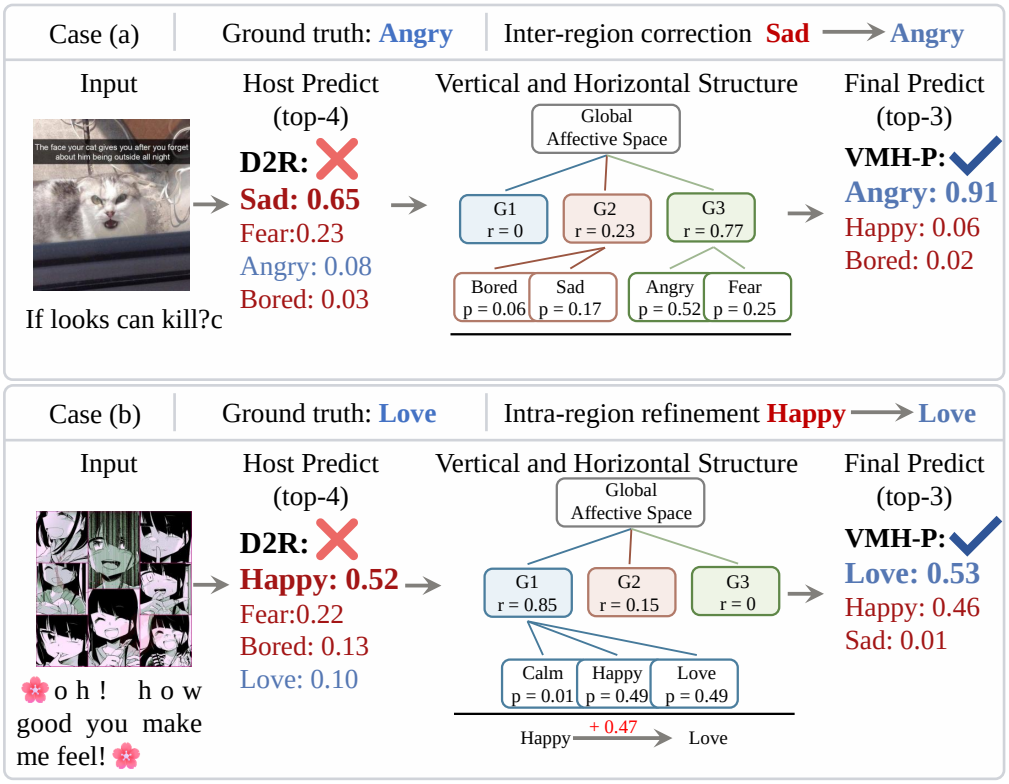


Figure 7: Case studies of VMH-P on TumEmo. Case (a) illustrates inter-region correction, while Case (b) shows intra-region refinement through horizontal relation propagation.