SIR-LMM: Learning Domain-Invariant Rationales for Out-of-Distribution Hateful Meme Detection

Yeqi Sun[1,2], Shizhong Peng[1,2], Yuqing Li[1,2], Huan Liu[*,1], Zheng Lin[1], Hao Cui[1,2], Huichuan Liu[3]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]State Grid Jiangsu Electric Power Co., Ltd
{sunyeqi, pengshizhong, liyuqing, liuhuan, linzheng, cuihao}@iie.ac.cn

***Abstract*—Explainable Hateful Meme Detection (EHMD) aims to classify multimodal content while providing natural language rationales, which is a critical capability for transparent content moderation. However, existing approaches typically operate under the independent and identically distributed (i.i.d.) assumption. Under real-world distribution shifts, Large Multimodal Models (LMMs) are prone to overfitting source-specific biases, such as recurring templates and visual styles, resulting in poor generalization on diverse domains. To address this, we propose SIR-LMM, a framework that learns a Sufficient Independent Representation (SIR) to decouple semantic hatefulness from domain variations. By imposing an information bottleneck, we conceptually partition the LMM into a feature extractor and a rationale generator. This structure is optimized via a two-stage paradigm: 1) conditional domain adversarial training on the extractor to ensure the representation is predictive of hatefulness (sufficiency) yet domain-invariant (conditional independence), and 2) fine-tuning the generator conditioned on the fixed SIR. This design ensures that high detection accuracy and rationale quality are maintained even under significant distribution shifts. Extensive experiments on three benchmarks (FHM, HarmC, and HarmP) demonstrate the robustness of our approach. Without requiring target domain rationale supervision, SIR-LMM outperforms state-of-the-art explainable baselines, improving average target-domain F1 score by 5.4% and BERTScore by 9.4%.**

***Index Terms*—Explainable Hateful Meme Detection, OOD Generalization, Representation Learning, Domain Adaptation**

I. Introduction

Internet memes have evolved from humorous cultural artifacts into vehicles for hate speech, utilizing the interplay of text and imagery to bypass unimodal filters [1]. Unlike conventional hate speech [2], detecting hateful memes requires disentangling complex multimodal interactions, where a benign image can become offensive when paired with specific text, and vice versa [3]. Given the subtle, context-dependent nature of such content, black-box predictions are insufficient for transparent content moderation. This necessitates Explainable Hateful Meme Detection (EHMD), which provides natural language rationales alongside predictions to foster trust and facilitate human-in-the-loop decision-making.

* Corresponding author.

This work was supported by the National Natural Science Foundation of China (Nos. 62472419, 62472420) and the Enterprise Project (No. E4V06811F3).

Disclaimer: *This paper includes examples of hateful content for academic analysis only. Such content does not reflect the authors' views.*

Existing EHMD approaches primarily rely on fine-tuning models with crowdsourced rationales [4] or prompting LMMs [5], [6]. While effective on standard benchmarks, these methods assume independent and identically distributed (i.i.d.) data. However, as shown in Fig. 1, real-world meme landscapes are dynamic, causing i.i.d.-trained models to exploit source-specific shortcuts (e.g., fonts) rather than semantic hatefulness. Consequently, these models suffer from degraded detection and explanation quality under distribution shifts.

Addressing this gap requires disentangling semantic features for hate detection from domain-specific noise. We identify three core challenges: (1) accurate detection across diverse distributions; (2) minimizing reliance on domain-specific biases; and (3) generating high-quality, domain-invariant rationales. To this end, we propose **SIR-LMM**, a framework that learns a *Sufficient Independent Representation* (SIR). By inserting an intermediate-layer information bottleneck, we conceptually partition the LMM into a feature extractor and a rationale generator, optimized via a corresponding two-stage paradigm. In the first stage, we establish the SIR via Conditional Domain Adversarial Training (CDAT). The intermediate representation is subjected to a minimax game involving a hate classifier and a domain discriminator: the extractor maximizes classification accuracy to ensure *sufficiency*, while simultaneously aiming to confuse the discriminator to enforce *conditional independence*. To mitigate negative transfer, the discriminator is conditioned on ground-truth labels, ensuring that alignment preserves class-discriminative semantics. In the second stage, we freeze the extractor and fine-tune upper layers as the generator, enabling the model to produce rationales grounded strictly in these robust, domain-invariant features.

We validate our approach through extensive experiments on three distinct datasets: FHM [1], HarmP [7], and HarmC [8]. Crucially, our framework utilizes rationale supervision only from the source domain (FHM), leveraging target-domain labels solely for invariance learning. Under this challenging setting, SIR-LMM achieves the best overall cross-domain performance among explainable baselines, yielding the top average target-domain results across both detection and explanation metrics. Moreover, it remains competitive with strong classification-only systems while additionally producing transferable rationales, making it better suited for practical, explanation-driven moderation under distribution shift.

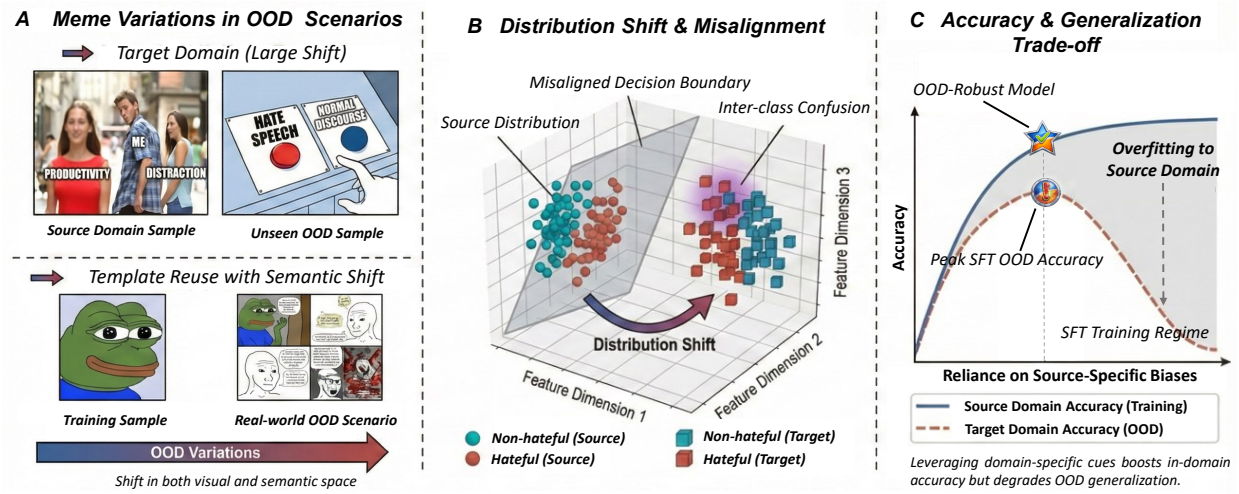

Fig. 1. Challenges of EHMD under domain shift. (A) Memes evolve across domains via template reuse, semantic drift, and emerging styles, producing OOD inputs relative to the source training set. (B) This shift causes feature-space misalignment; a source-trained decision boundary can yield inter-class confusion on the target domain. (C) In-domain fine-tuning often exploits source-specific shortcuts, improving source accuracy but hurting OOD generalization, motivating domain-invariant representations for robust detection and explanations.

Our contributions are summarized as follows:

- We identify the limitations of existing EHMD methods under distribution shifts and propose a robust framework, SIR-LMM, to address the challenge of supervised domain generalization in real-world moderation.
- We introduce a mechanism that integrates Conditional Domain Adversarial Training with parameter-efficient fine-tuning, allowing the model to learn a Sufficient Independent Representation that filters out domain-specific spurious correlations.
- Extensive experiments demonstrate that SIR-LMM significantly outperforms baselines in cross-domain settings, generating high-quality rationales without any target-domain rationale supervision.

The code is publicly available at https://github.com/Starfulllll/SIR-LMM.

II. Related Work

A. Hateful Meme Detection and Explainability

The introduction of the Hateful Memes Challenge [1] has spurred significant research into multimodal detection. Early approaches focused on optimizing cross-modal feature fusion [9]–[11] or integrating external knowledge [12]–[14] to capture implicit context. Recently, the field has shifted towards LMMs, utilizing instruction tuning [15], [16] and Chain-of-Thought (CoT) reasoning [17]–[19] to enhance comprehension. To improve transparency, EHMD has emerged and has been studied via either rationale-based fine-tuning [4] or zero-shot prompting strategies [5], [6], [20], [21]. However, these methods typically operate under the assumption that training and testing data are i.i.d. As noted by Aggarwal et al. [22], such models often overfit to source-domain shortcuts (e.g., specific templates or styles) rather than learning semantic hatefulness, leading to poor generalization and inconsistent explanations when facing real-world distribution shifts.

B. Domain Adaptation and Generalization

Domain Adaptation (DA) and Generalization (DG) address distribution shifts by learning transferable features. A seminal approach, DANN [23], employs adversarial training to align marginal distributions, with conditional variants [24] further incorporating class information to prevent negative transfer. Other strategies include invariant risk minimization [25], statistical discrepancy minimization [26], [27], and contrastive learning [28], [29]. However, transferring these discriminative paradigms to EHMD is non-trivial, as standard alignment often neglects the generative nature of LMMs, risking the collapse of linguistic coherence required for explanation. SIR-LMM addresses this by conditional adversarial alignment to an intermediate representational bottleneck, ensuring the learned features are sufficient for detection, invariant to domain shifts, and compatible with coherent rationale generation.

III. Methodology

We introduce SIR-LMM to improve robustness under domain shift by separating hate-related semantics from domain-specific artifacts. The core idea is to impose an intermediate-layer bottleneck in the LMM and learn a *Sufficient Independent Representation* (SIR) that captures label-relevant information while filtering domain cues. As shown in Fig. 2, training proceeds in two steps: (1) learning the SIR via CDAT, and (2) freezing the extractor and fine-tuning the upper layers to generate *formatted* rationales, which jointly provide a rationale and an explicit label prediction for each input.

A. Problem Formulation

Let $\mathcal{X}$, $\mathcal{Y} = \{0, 1\}$, and $\mathcal{R}$ denote the multimodal input, label, and rationale spaces, respectively. We operate in a supervised domain adaptation setting with a source domain $\mathcal{S} = \{(x_i^s, y_i^s, r_i^s)\}_{i=1}^{N_s}$ and a target domain $\mathcal{T} = \{(x_j^t, y_j^t)\}_{j=1}^{N_t}$, where rationales are exclusive to $\mathcal{S}$. Domains follow distinct

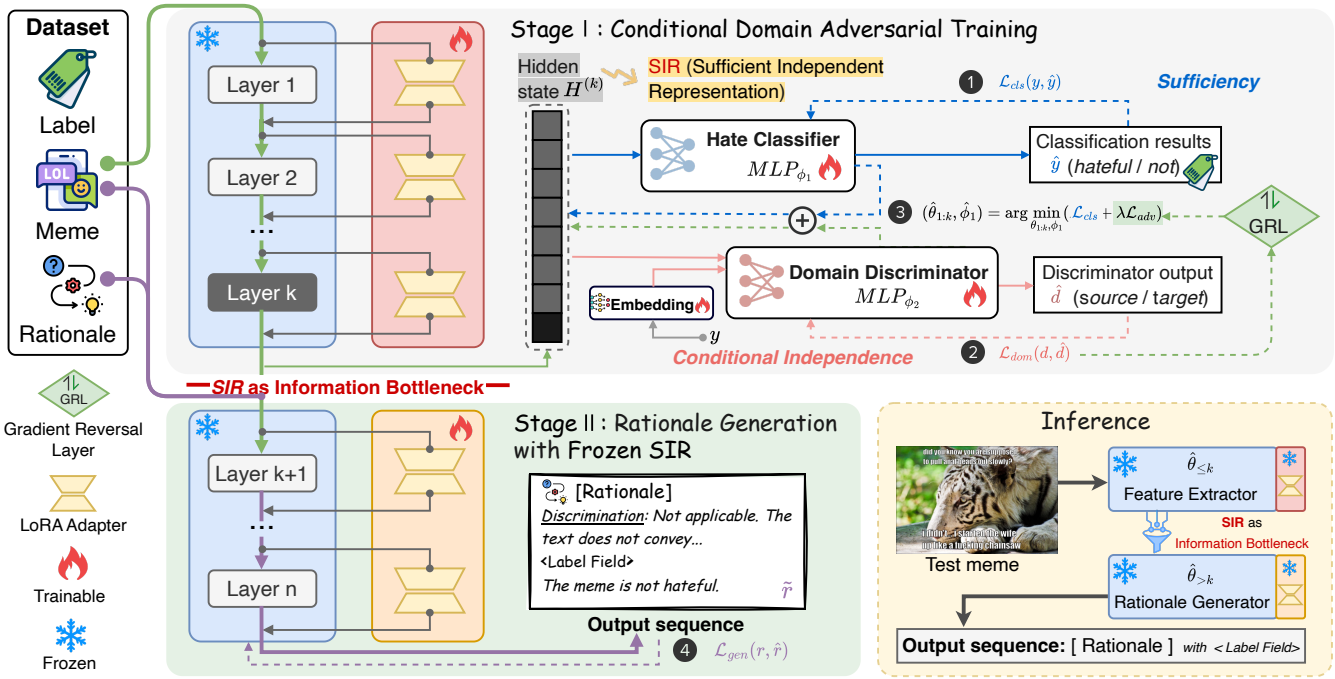

Fig. 2. Overview of **SIR-LMM**. The LMM is conceptually partitioned into a feature extractor (Layers $1 \sim k$) and a rationale generator (Layers $k{+}1 \sim n$) via an intermediate information bottleneck. **Stage I:** The extractor is optimized via conditional domain adversarial training. A hate classifier ensures the learned *Sufficient Independent Representation* (SIR) retains semantic sufficiency, while a label-conditioned domain discriminator enforces conditional independence via a Gradient Reversal Layer (GRL). **Stage II:** The SIR extractor is frozen to prevent the reintroduction of source-specific biases, and the upper layers are fine-tuned to generate formatted rationales. **Inference:** The model utilizes the robust SIR to generate domain-invariant explanations and predictions.

marginal distributions $P_{\mathcal{S}}(x) \neq P_{\mathcal{T}}(x)$, indexed by a domain indicator $d \in \{0, 1\}$ (0 for $\mathcal{S}$, 1 for $\mathcal{T}$). Our objective is to learn a function $f_\Theta : \mathcal{X} \to \mathcal{Y} \times \mathcal{R}$ that generalizes to $\mathcal{T}$ by leveraging source-domain supervision.

B. Theoretical Framework

We adopt a decoder-only LMM backbone parameterized by Θ. To ensure parameter efficiency, we employ Low-Rank Adaptation (LoRA) [30]. The LMM is conceptually partitioned into a feature extractor $E_{\theta_{\leq k}}$ (layers 1 to k) and a generator $G_{\theta_{>k}}$ (layers $k+1$ to N_{layer}). We set k to an intermediate layer (i.e., roughly the middle of the network) to obtain a representation that is expressive yet amenable to invariance learning. This choice is evaluated in ablation studies.

Given an input meme x, the extractor produces the layer-k hidden-state sequence $H^{(k)} = E_{\theta_{\leq k}}(x) \in \mathbb{R}^{T \times h}$ for a sequence length T. We define the SIR vector as the hidden state of the last token at layer k:

$$\mathbf{z} = H_T^{(k)} \in \mathbb{R}^h. \tag{1}$$

To learn $\mathbf{z}$, we introduce two modules used in Stage I: (i) a hate classifier C_ϕ and (ii) a domain discriminator D_ψ.

For robust OOD detection, $\mathbf{z}$ should satisfy two information-theoretic properties:

1) *Sufficiency:* $\mathbf{z}$ must retain maximal information about y, i.e., maximizing $I(\mathbf{z}; y)$.
2) *Conditional Independence:* $\mathbf{z}$ must be invariant to domain d given y, i.e., minimizing $I(\mathbf{z}; d \mid y)$.

Consequently, we seek to optimize:

$$\max_{\theta_{\leq k}} \quad I(\mathbf{z}; y) - \lambda I(\mathbf{z}; d \mid y), \tag{2}$$

where λ is a hyperparameter governing the trade-off between semantic sufficiency and domain invariance. Since exact mutual information computation is intractable, we approximate Eq. (2) with a variational adversarial objective (Stage I) expressed as a minimax game.

C. Stage I: Conditional Domain Adversarial Training

This stage focuses on optimizing the extractor parameters $\theta_{\leq k}$ to approximate Eq. (2). We train C_ϕ to learn label-predictive features, while adversarially training $E_{\theta_{\leq k}}$ against D_ψ to remove domain information conditioned on class.

1) Maximizing Sufficiency: To ensure $\mathbf{z}$ is predictive of hatefulness, we train the classifier C_ϕ to minimize the cross-entropy loss on labeled samples from both domains:

$$\mathcal{L}_{\text{cls}}(\theta_{\leq k}, \phi) = \mathbb{E}_{(x,y) \sim \mathcal{S} \cup \mathcal{T}} \left[-\log P(y \mid \mathbf{z}; \phi) \right], \tag{3}$$

where $P(\cdot \mid \mathbf{z}; \phi) = \text{softmax}(C_\phi(\mathbf{z}))$ denotes the predicted class distribution. Minimizing $\mathcal{L}_{\text{cls}}$ is equivalent to maximizing a variational lower bound on $I(\mathbf{z}; y)$.

2) Enforcing Conditional Independence: Standard DA aligns marginal distributions $P(\mathbf{z})$, which risks misaligning class-conditional structures, such as mapping hateful source memes to non-hateful target memes. Instead, we employ

CDAT to align $P(\mathbf{z} \mid y)$. Specifically, we condition the domain discriminator on the ground-truth class label:

$$\mathbf{e}_y = \mathrm{Embed}(y), \qquad \mathbf{v} = \mathbf{z} \oplus \mathbf{e}_y, \tag{4}$$

where $\oplus$ denotes concatenation and $\mathrm{Embed}(\cdot)$ is a learnable embedding layer that maps y to a dense representation. This conditioning ensures that alignment only occurs between samples of the same class across domains, effectively mitigating stylistic noise while retaining class-discriminative semantics.

The discriminator D_ψ aims to identify the domain d from $\mathbf{v}$. The domain classification loss is defined as:

$$\mathcal{L}_{\text{dom}}(\theta_{\le k}, \psi) = \mathbb{E}_{(x,y,d)\sim\mathcal{S}\cup\mathcal{T}} \left[\mathcal{L}_{\text{BCE}}(D_\psi(\mathbf{v}), d)\right]. \tag{5}$$

Combining the classification and domain objectives, Stage I is formulated as a minimax game:

$$\min_{\theta_{\le k},\,\phi} \max_{\psi} \quad \mathcal{L}_{\text{cls}}(\theta_{\le k}, \phi) \;-\; \lambda\, \mathcal{L}_{\text{dom}}(\theta_{\le k}, \psi). \tag{6}$$

Maximizing w.r.t. ψ reduces $\mathcal{L}_{\text{dom}}$, while minimizing w.r.t. $\theta_{\le k}$ increases $\mathcal{L}_{\text{dom}}$, making domains hard to predict.

In practice, we realize this objective via a Gradient Reversal Layer (GRL) [23]. Define the adversarial loss as:

$$\mathcal{L}_{\text{adv}}(\theta_{\le k}, \psi) = \mathbb{E}_{(x,y,d)\sim\mathcal{S}\cup\mathcal{T}} \left[\mathcal{L}_{\text{BCE}}(D_\psi(\mathrm{GRL}(\mathbf{v})), d)\right], \tag{7}$$

where $\mathrm{GRL}(\cdot)$ acts as the identity during forward propagation but reverses the gradient sign during backpropagation. This enables end-to-end training via a unified minimization:

$$\min_{\theta_{\le k},\,\phi,\,\psi} \quad \mathcal{L}_{\text{cls}}(\theta_{\le k}, \phi) \;+\; \lambda\, \mathcal{L}_{\text{adv}}(\theta_{\le k}, \psi). \tag{8}$$

This formulation serves as an approximation to Eq. (2). The optimal parameters $\hat{\theta}_{\le k}$ are obtained upon convergence.

D. Stage II: Rationale Generation with Frozen SIR

In the second stage, we train the upper layers to generate formatted rationales. A critical insight of our method is that fine-tuning the entire model on source rationales would reintroduce source-specific biases. Therefore, we freeze the extractor output at layer k as:

$$H^{(k)}_{\text{fixed}} = \mathrm{sg}\Big(E_{\hat{\theta}_{\le k}}(x)\Big), \tag{9}$$

where $\mathrm{sg}(\cdot)$ denotes the *stop-gradient* operator. We then optimize the parameters $\theta_{>k}$ of the generator $G_{\theta_{>k}}$.

To enable the deployed system to return both a label and an explanation through a single generation, we train the generator to output a sequence that contains natural language reasoning and ends with a *label field*. Denote this formatted sequence as $\tilde{r} = (\tilde{w}_1, \ldots, \tilde{w}_{\tilde{L}})$. Using source data only, the objective is:

$$\mathcal{L}_{\text{gen}}(\theta_{>k}) = \mathbb{E}_{(x,\tilde{r})\sim\mathcal{S}} \left[-\sum_{t=1}^{\tilde{L}} \log G_{\theta_{>k}}(\tilde{w}_t \mid \tilde{w}_{<t}, H^{(k)}_{\text{fixed}}) \right]. \tag{10}$$

This ensures the reasoning process is grounded in the domain-invariant semantics captured by the frozen intermediate representation. Upon convergence, we obtain the optimal upper-layer parameters $\hat{\theta}_{>k}$.

E. Inference

During inference on a test meme x^*, the model performs a forward pass using the parameters $\hat{\Theta} = \{\hat{\theta}_{\le k}, \hat{\theta}_{>k}\}$. The frozen lower layers compute the intermediate representation $H^{(k)}(x^*)$, and the generator $G_{\hat{\theta}_{>k}}$ autoregressively produces the output sequence $\hat{r}$ conditioned on this representation. We regard the entire sequence $\hat{r}$ as the natural language rationale. To perform classification, we extract the predicted label $\hat{y}$ by parsing the formatted *label field* contained within $\hat{r}$.

IV. Experiments

A. Experimental Setup

1) Datasets and Evaluation Metrics: We conduct experiments on three public benchmarks for EHMD: **FHM** [1], **HarmC** [7], and **HarmP** [8]. FHM is deliberately constructed to stress multimodal reasoning by including challenging counterfactual-style cases where the image or the overlaid text is minimally edited to flip the label while preserving superficial cues. HarmC and HarmP focus on COVID-19 political content and U.S. politics, respectively. All three datasets only provide classification labels and do not include human-written explanations. To enable explainable HMD evaluation, we adopt the rationales released in MemeMind [31] as reference explanations for FHM, HarmC, and HarmP. These rationales serve as reference targets for training and evaluating rationale generation, and allow us to quantify explanation quality with standard automatic metrics. In addition, to ensure a consistent experimental protocol, we follow MemeMind for the training/test splits of all three datasets. Detailed statistics of the datasets are summarized in Table II. We evaluate classification performance using Accuracy and Macro-F1, consistent with prior benchmarks [4], [32]. For rationale generation, we employ METEOR [33] and BERTScore [34] to assess lexical alignment and semantic consistency, respectively.

2) Implementation Details: Our proposed SIR-LMM is implemented using the PyTorch framework and trained on a system equipped with 8 NVIDIA A800 GPUs. We adopt LLaVA-OneVision-7B (LLaVA-OV) [35] as the LMM backbone. To achieve parameter-efficient fine-tuning, we apply LoRA [30] to all linear layers of the backbone, with the rank and scaling factor set to $r = 64$ and $\alpha = 128$, respectively. The model is optimized using the AdamW optimizer with a learning rate of 3×10^{-5} and $\beta_1 = 0.9, \beta_2 = 0.999$. The global batch size is set to 40. Regarding the method-specific hyperparameters, the intermediate layer index for SIR extraction is set to $k = 16$. Regarding the auxiliary modules in Stage I, both the hate classifier and the domain discriminator are implemented as Multi-Layer Perceptrons (MLPs) with a hidden dimension of 512. Additionally, the label embedding used for conditional alignment utilizes a dimension of 32. The conditional domain adversarial weight is set to $\lambda = 0.5$. The training procedure consists of a two-stage pipeline: 3 epochs for domain-invariant representation learning (Stage I), followed by an epoch for rationale generation fine-tuning (Stage II).

TABLE I

PERFORMANCE COMPARISON ON SOURCE DOMAIN (FHM) AND TARGET DOMAINS (HARMC, HARMP). **C-METRICS** DENOTES CLASSIFICATION PERFORMANCE (ACC, F1), AND **E-METRICS** DENOTES EXPLANATION QUALITY (METEOR, BERTSCORE). THE "AVG." COLUMN REPORTS THE AVERAGE PERFORMANCE ACROSS THE TWO TARGET DOMAINS.

Model	FHM (Source)				HarmC (Target)				HarmP (Target)				Target Avg.			
	C-Metrics		E-Metrics		C-Metrics		E-Metrics		C-Metrics		E-Metrics		C-Metrics		E-Metrics	
	Acc	F1	MET	BS	Acc	F1	MET	BS	Acc	F1	MET	BS	Acc	F1	MET	BS
Classification Baselines (Rationale Unavailable)																
LLaVA-OV w/ 0-shot (7B)	76.7	81.7	–	–	59.9	55.2	–	–	70.7	72.8	–	–	65.3	64.0	–	–
Qwen2.5-VL w/ 0-shot (32B)	72.8	72.6	–	–	52.4	49.1	–	–	69.7	64.7	–	–	61.1	56.9	–	–
HateCLIPper	78.2	83.9	–	–	83.9	**86.6**	–	–	77.1	74.3	–	–	80.5	80.5	–	–
Explainable Methods																
ExplainHM (7B)	77.6	70.3	18.4	63.7	**84.3**	80.7	18.3	63.0	73.4	69.6	19.0	63.8	78.9	75.2	18.7	63.4
LLaVA-OV w/ CoT (7B)	66.7	80.0	29.8	71.6	59.9	74.9	33.0	73.0	45.1	62.2	33.7	73.8	52.5	68.6	33.4	73.4
Qwen2.5-VL w/ CoT (32B)	**80.8**	84.7	39.9	75.1	68.0	68.4	34.5	72.9	73.4	71.8	33.4	73.0	70.7	70.1	34.0	73.0
LLaVA-OV w/ SFT (7B)	79.3	83.4	48.1	80.5	76.1	76.8	49.9	81.6	76.4	73.7	48.6	81.5	76.3	75.3	49.3	81.6
LoraHub (7B)	79.8	84.3	48.2	80.4	74.5	76.0	**51.9**	81.6	73.3	69.1	49.2	81.0	73.9	72.6	50.6	81.3
SIR-LMM (7B)	80.6	**84.8**	**48.6**	**90.3**	82.7	84.6	51.5	**90.9**	**78.5**	**76.7**	**50.4**	**91.0**	**80.6**	**80.7**	**51.0**	**91.0**

TABLE II

STATISTICS OF THE THREE BENCHMARK DATASETS.

Dataset	Train		Test	
	Hateful	Not Hateful	Hateful	Not Hateful
FHM	3,796	1,874	1,717	859
HarmC	1,487	1,006	626	417
HarmP	962	1,210	413	503

3) Baselines: We compare SIR-LMM against a diverse set of methods categorized into two paradigms: (i) **Classification Baselines**, which focus solely on binary detection. We evaluate **HateCLIPper** [9], a specialized contrastive learning framework, alongside state-of-the-art LMMs—specifically **LLaVA-OV** [35] and **Qwen2.5-VL** [36]—under a zero-shot prompting setting. (ii) **Explainable Baselines**, which provide reasoning alongside predictions. We employ **CoT** prompting [37] on the aforementioned LMMs to elicit reasoning. For supervised approaches, we compare against **ExplainHM** [6], a prior state-of-the-art generative HMD model, and **LLaVA-SFT**, which applies LoRA fine-tuning on source rationales. Furthermore, to evaluate existing OOD generalization methods, we adapt **LoraHub** [38] to EHMD, which was originally designed for text generation via module composition.

B. Main Results

Table I summarizes the results on the source domain and two target domains. Overall, SIR-LMM delivers the strongest cross-domain performance among all explainable methods, despite using a 7B backbone. On the *Target Avg.*, SIR-LMM achieves the best results on all four metrics, outperforming both prompting-based LMM baselines and rationale-supervised fine-tuning baselines. In particular, we highlight the comparison with LoraHub, which we adapted to strictly follow the supervised domain adaptation setting. Despite utilizing target-domain supervision similar to our approach, LoraHub yields significantly lower average classification and generation scores on target domains. This performance gap validates the effectiveness of SIR-LMM, demonstrating that explicit invariance learning is far more robust against distribution shifts than simple module composition. Notably, this advantage holds even against Qwen2.5-VL (32B) under CoT prompting, indicating that robustness here is driven more by representation learning than by model scale. On the source domain, SIR-LMM attains the best F1 and the strongest explanation quality, while remaining competitive in accuracy.

We also observe that label-only classification baselines appear competitive on target sets, which we attribute to the nature of the supervised adaptation protocol. First, the availability of target labels during training effectively reduces the task for these models to a standard i.i.d. supervised setting, allowing them to bypass the challenge of distribution shift. Second, optimizing solely for labels encourages exploiting domain-specific shortcuts (e.g., recurring templates) [22], yielding high test accuracy that masks a lack of robust reasoning. Crucially, these baselines cannot generate rationales, rendering them insufficient for transparent and explainable moderation.

Finally, we find that advanced LMM CoT prompting is still insufficient for EHMD across domains, suggesting that hateful meme phenomena are underrepresented or weakly grounded in generic pretraining corpora. Supervised rationale fine-tuning substantially improves in-domain explanation quality and yields source-domain performance close to ours, supporting the necessity of explicitly learning hate-related representations. However, the gap between these fine-tuned baselines and SIR-LMM widens on target domains across all four averaged metrics, consistent with the hypothesis that

TABLE III

ABLATION STUDY OF SIR-LMM. WE REPORT THE PERFORMANCE ON THE SOURCE AND TARGET DOMAINS WHEN REMOVING SPECIFIC COMPONENTS. "*w/o*" DENOTES THE REMOVAL OF A MODULE. *Cond. DA*, *Hate Head*, AND *Domain Disc.* REFER TO THE CONDITIONAL DOMAIN ADAPTATION MECHANISM, THE HATE CLASSIFICATION HEAD, AND THE DOMAIN DISCRIMINATOR, RESPECTIVELY.

Variant	FHM (Source)				HarmC (Target)				HarmP (Target)				Target Avg.			
	C-Metrics		E-Metrics		C-Metrics		E-Metrics		C-Metrics		E-Metrics		C-Metrics		E-Metrics	
	Acc	F1	MET	BS	Acc	F1	MET	BS	Acc	F1	MET	BS	Acc	F1	MET	BS
SIR-LMM	80.6	84.8	48.6	90.3	82.7	84.6	51.5	90.9	78.5	76.7	50.4	91.0	80.6	80.7	51.0	91.0
w/o Cond. DA	79.6	84.4	47.5	90.3	82.5	85.2	49.4	90.7	74.0	73.4	48.6	90.9	78.3	79.3	49.0	90.8
w/o Domain Disc.	79.4	83.6	46.2	80.1	77.6	78.7	46.7	80.8	73.5	69.4	46.2	81.0	75.6	74.1	46.5	80.9
w/o Hate Head	68.2	72.1	45.1	79.2	60.0	75.0	25.9	70.0	45.4	62.3	25.5	70.3	52.7	68.7	25.7	70.2

end-to-end rationale tuning can reintroduce source-specific biases. In contrast, by explicitly learning a domain-invariant intermediate representation and freezing it for generation, SIR-LMM maintains stronger detection performance and produces more stable, transferable rationales under domain shift.

C. Analysis

1) Ablation Study: To evaluate the contribution of each component within the SIR-LMM framework and verify the necessity of the proposed mechanisms, we conduct an ablation study by independently removing key modules. Specifically, we investigate three variants: (i) *w/o Cond. DA*: The domain discriminator is trained to classify domains based solely on the SIR $\mathbf{z}$, without conditioning on the class label. This reverts the method to standard marginal distribution alignment. (ii) *w/o Domain Disc.*: The domain discriminator is removed entirely. This leaves Stage I to focus solely on source-domain classification, thereby quantifying the transferability gain directly attributed to adversarial adaptation. (iii) *w/o Hate Head*: The hate classifier is removed. Stage I focuses exclusively on confusing the domain discriminator, relying entirely on Stage II for task-specific supervision.

The detailed results are presented in Table III. We summarize our observations as follows: ***(i) Necessity of conditional alignment.*** Comparing the full SIR-LMM with *w/o Cond. DA*, we observe that removing the class conditioning leads to a notable decline in generalization performance, despite maintaining competitive results on the source domain. Specifically, on the HarmP dataset, the classification accuracy drops by 4.5% and the METEOR score decreases by 1.8%. This degradation indicates that aligning marginal distributions indiscriminately forces the model to map hateful samples from the source domain to non-hateful samples in the target domain (or vice versa) to satisfy the invariance constraint. The conditional mechanism in SIR-LMM effectively mitigates this negative transfer by preserving class-discriminative structures during alignment. ***(ii) Impact of domain adversarial training.*** The removal of the domain discriminator (*w/o Domain Disc.*) results in a significant reduction in model robustness. Without the adversarial constraint, the intermediate representation $\mathbf{z}$ overfits to source-specific biases (e.g., visual styles) captured by the auxiliary classifier. Consequently, the frozen representation passed to the generator in Stage II lacks the necessary domain invariance. As shown in Table III, the average Target F1 score drops by 6.6%, and the BERTScore plummets by over 10 points compared to the full model. Interestingly, the performance of this variant in OOD settings is inferior to standard fine-tuning baselines, suggesting that a representation optimized solely for a linear classification head on source data may create a "feature gap", rendering it suboptimal for the decoder-only generator when domain shifts occur. ***(iii) Role of task-specific supervision.*** Finally, the *w/o Hate Head* variant exhibits the poorest performance across all metrics. Without the hate classification objective in Stage I, the representation learning is unguided regarding the semantic task, yielding a SIR that is neither predictive nor meaningful. This confirms that the hate classification head is fundamental for anchoring semantic content in the intermediate representation. In summary, all three components are indispensable. SIR-LMM effectively integrates them to learn a representation that is both semantically sufficient and domain-independent.

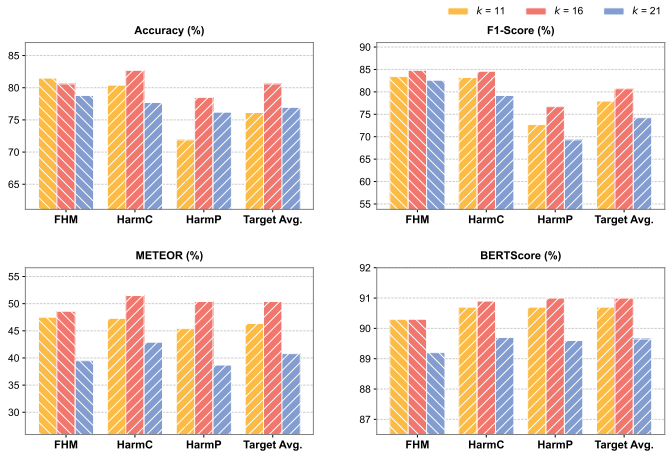

Fig. 3. Impact of the intermediate layer index k on model performance across the source (FHM) and target (HarmC, HarmP) domains. The results demonstrate that extracting the SIR from intermediate layers (e.g., $k = 16$) yields the optimal balance between task performance and OOD robustness. The model exhibits stability, maintaining high performance even with moderate variations in layer depth.

2) Impact of Layer Depth on Representation Robustness: To determine the optimal depth for extracting the SIR, we analyze the sensitivity of SIR-LMM to the intermediate layer

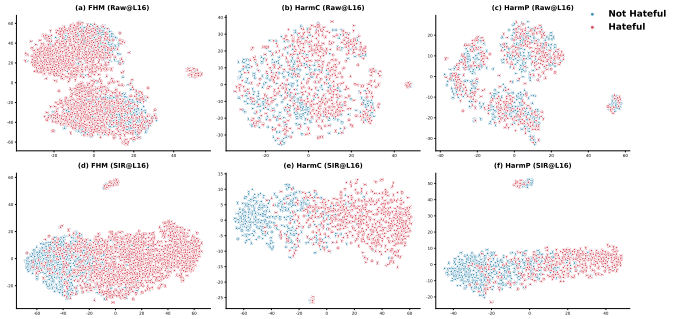

Fig. 4. t-SNE visualization of feature distributions colored by class labels. Row 1: raw backbone features show a disordered mixture of classes. Row 2: the learned SIR exhibits enhanced cluster purity and improved separability between hateful and non-hateful samples.

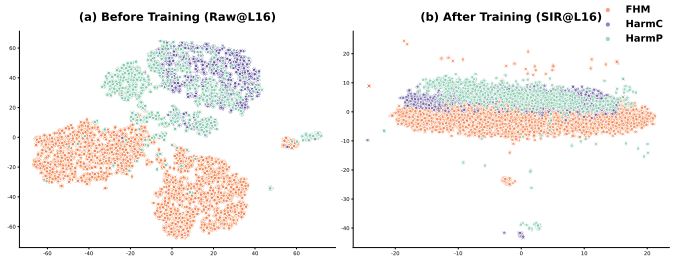

Fig. 5. t-SNE visualization of feature distributions colored by domain. (a): raw representations appear as divergent, disjoint clusters. (b): the SIR features display consistent manifold shapes and geometric alignment across datasets, indicating reduced domain discrepancy.

index k. Fig. 3 presents the performance variations across classification and rationale generation metrics when k is varied among $\{11, 16, 21\}$ within the LMM backbone. We observe that extracting the SIR from the intermediate layers consistently yields superior results compared to shallower or deeper layers. Notably, the intermediate layer achieves the highest scores in both source-domain accuracy and target-domain generalization, while maximizing explanation quality metrics. We attribute this phenomenon to two primary factors:

(i) Semantic abstraction hierarchy. Recent findings in representation engineering suggest that Large Language Models process information hierarchically [39], [40]. Shallow layers encode modality-specific surface features (e.g., visual edges or lexical syntax), which are domain-dependent and prone to spurious correlations, while the deepest layers focus on next-token distributions rather than holistic semantic understanding. Intermediate layers function as a "semantic hub", capturing abstract concepts essential for EHMD while remaining amenable to the domain-invariance constraints of our CDAT mechanism.

(ii) Parameter allocation equilibrium. The choice of k dictates the partition of learnable parameters between the Feature Extractor ($E_{\theta_{\leq k}}$) and the Rationale Generator ($G_{\theta_{>k}}$). Setting k too low restricts the Extractor's capacity to learn a truly domain-invariant representation, leading to "leakage" of domain styles into the SIR. Conversely, setting k too high leaves the Generator with insufficient capacity to translate the SIR into coherent natural language explanations. An intermediate k establishes a Nash Equilibrium among the competing objectives of hate classification, adversarial robustness, and rationale generation, ensuring that the learned representation is both discriminative and linguistically decipherable.

Furthermore, empirical results indicate that SIR-LMM is not sensitive to the exact choice of k. As illustrated in Fig. 3, performance fluctuations remain within a narrow margin when shifting the extraction point by several layers. This stability suggests that the proposed framework is robust and does not require extensive hyperparameter tuning to achieve generalization in real-world scenarios.

3) Visualization of Sufficient Independent Representation: To intuitively assess the quality of the learned representations, we visualize the feature distributions of the test sets from FHM, HarmC, and HarmP using t-SNE. We compare the raw intermediate features from the frozen backbone (Layer 16) against our optimized SIR. (i) Semantic sufficiency. Fig. 4 illustrates the sample distributions colored by ground-truth labels. In the raw feature space (Row 1), hateful and non-hateful samples are heavily entangled, indicating that the pre-trained backbone does not spontaneously isolate hate-specific semantics at this depth. In contrast, the SIR (Row 2) exhibits significantly enhanced separability. While t-SNE projections do not imply linear decision boundaries, the SIR space reveals distinct clustering patterns where samples of the same class aggregate more cohesively. This suggests that the auxiliary classification objective successfully injects semantic discriminability into the representation, facilitating the disentanglement of hateful concepts from background noise. (ii) Domain independence. Fig. 5 depicts the distributions colored by dataset membership. The raw representations (a) form disjoint, scattered clusters, reflecting the significant domain gap caused by varying collection sources and styles. Conversely, the SIR distributions (b) demonstrate a high degree of manifold alignment. Although the three datasets do not perfectly overlap, likely due to intrinsic content differences, their projections exhibit highly consistent geometric structures and shapes. This topological similarity indicates that the CDAT has effectively mitigated the distributional divergence, forcing the model to adopt a domain-invariant encoding scheme that generalizes across the source and target domains.

V. Conclusion

We propose SIR-LMM to disentangle semantic hatefulness from domain-specific noise via conditional domain adversarial training. By enforcing a Sufficient Independent Representation (SIR) at an intermediate bottleneck, our framework effectively filters stylistic artifacts while preserving class-discriminative semantics. Experiments across three benchmarks demonstrate that SIR-LMM significantly improves both detection accuracy and rationale quality on target domains, highlighting the necessity of explicit invariance learning in dynamic meme environments. Future work will explore extending this paradigm to zero-shot adaptation and investigating scalability across varying LMM architectures.

REFERENCES

[1] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Advances in neural information processing systems*, vol. 33, pp. 2611–2624, 2020.

[2] F. Alkomah and X. Ma, "A literature review of textual hate speech detection methods and datasets," *Information*, vol. 13, no. 6, 2022.

[3] S. Sharma, F. Alam, M. S. Akhtar, D. Dimitrov, G. D. S. Martino, H. Firooz, A. Halevy, F. Silvestri, P. Nakov, and T. Chakraborty, "Detecting and understanding harmful memes: A survey," *arXiv preprint arXiv:2205.04274*, 2022.

[4] M. B. Kmainasi, A. Hasnat, M. A. Hasan, A. E. Shahroor, and F. Alam, "Memeintel: Explainable detection of propagandistic and hateful memes," *arXiv preprint arXiv:2502.16612*, 2025.

[5] M. S. Hee and R. K.-W. Lee, "Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, 2025, pp. 774–785.

[6] H. Lin, Z. Luo, W. Gao, J. Ma, B. Wang, and R. Yang, "Towards explainable harmful meme detection through multimodal debate between large language models," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 2359–2370.

[7] S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M. S. Akhtar, P. Nakov, and T. Chakraborty, "Detecting harmful memes and their targets," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2783–2796.

[8] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty, "MOMENTA: A multimodal framework for detecting harmful memes and their targets," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 4439–4455.

[9] G. K. K. K. Nandakumar, "Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features," *NLP4PI 2022*, p. 171, 2022.

[10] R. K.-W. Lee, R. Cao, Z. Fan, J. Jiang, and W.-H. Chong, "Disentangling hate in online memes," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 5138–5147.

[11] J. Mei, J. Chen, W. Lin, B. Byrne, and M. Tomalin, "Improving hateful meme detection through retrieval-guided contrastive learning," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 5333–5347.

[12] B. Grasso, V. La Gatta, V. Moscato, and G. Sperlì, "Kermit: Knowledge-empowered model in harmful meme detection," *Information Fusion*, vol. 106, p. 102269, 2024.

[13] M. Tzelepi and V. Mezaris, "Improving multimodal hateful meme detection exploiting lmm-generated knowledge," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 202–211.

[14] R. Cao, M. S. Hee, A. Kuek, W.-H. Chong, R. K.-W. Lee, and J. Jiang, "Pro-cap: Leveraging a frozen vision-language model for hateful meme detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5244–5252.

[15] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 49 250–49 267, 2023.

[16] Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Y. Wang, Y. Rao, J. Liu, T. Huang, and X. Wang, "Generative multimodal models are in-context learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 398–14 409.

[17] F. Pan, A. T. Luu, and X. Wu, "Detecting harmful memes with decoupled understanding and guided cot reasoning," *arXiv preprint arXiv:2506.08477*, 2025.

[18] Y. Hu, O. Stretcu, C.-T. Lu, K. Viswanathan, K. Hata, E. Luo, R. Krishna, and A. Fuxman, "Visual program distillation: Distilling tools and programmatic reasoning into vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9590–9601.

[19] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[20] R. Cao, R. K.-W. Lee, W.-H. Chong, and J. Jiang, "Prompting for multimodal hateful meme classification," in *Proceedings of the 2022 conference on empirical methods in natural language processing*, 2022, pp. 321–332.

[21] M. S. Hee, R. K.-W. Lee, and W.-H. Chong, "On explaining multimodal hateful meme detection models," in *Proceedings of the ACM web conference 2022*, 2022, pp. 3651–3655.

[22] P. Aggarwal, J. Mehrabanian, W. Huang, Ö. Alaçam, and T. Zesch, "Text or image? what is more important in cross-domain generalization capabilities of hate meme detection models?" in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 104–117.

[23] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[24] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *Advances in neural information processing systems*, vol. 31, 2018.

[25] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[26] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The journal of machine learning research*, vol. 13, no. 1, pp. 723–773, 2012.

[27] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[28] X. Yao, Y. Bai, X. Zhang, Y. Zhang, Q. Sun, R. Chen, R. Li, and B. Yu, "Pcl: Proxy-based contrastive learning for domain generalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7097–7107.

[29] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *ICLR*, 2021.

[30] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[31] H. Gu, Q. Yu, S. Hou, Z. Fang, H. Wu, and Z. He, "Mememind: A large-scale multimodal dataset with chain-of-thought reasoning for harmful meme detection," *arXiv preprint arXiv:2506.18919*, 2025.

[32] H. Lin, Z. Luo, W. Gao, J. Ma, B. Wang, and R. Yang, "Towards explainable harmful meme detection through multimodal debate between large language models," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 2359–2370.

[33] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[34] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[35] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li, "LLaVA-onevision: Easy visual task transfer," *Transactions on Machine Learning Research*, 2025.

[36] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *CoRR*, 2025.

[37] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[38] C. Huang, Q. Liu, B. Y. Lin, T. Pang, C. Du, and M. Lin, "Lorahub: Efficient cross-task generalization via dynamic lora composition," *CoRR*, 2023.

[39] S. Marks and M. Tegmark, "The geometry of truth: Emergent linear structure in large language model representations of true/false datasets," in *First Conference on Language Modeling*, 2024.

[40] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski *et al.*, "Representation engineering: A top-down approach to ai transparency," *CoRR*, 2023.