

TCHV: Type-Conditioned Hypothesis Verification for Multimodal Named Entity Recognition

Feiyu Han^{1,2}, Yang Liu², Yongxiu Xu^{3,*}, Yixiao Jiang^{1,2}, Zhaoqi Zhou⁴, Chuan Zhou^{2,5,*}

¹School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing, China

²Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

³Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

⁴Huawei Technologies Co., Ltd., Beijing, China

⁵School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

*Corresponding authors: Yongxiu Xu and Chuan Zhou

{hanfeiyu, liuyang, jiangyixiao, zhouchuan}@amss.ac.cn

xuyongxiu@iie.ac.cn, zhouzhaoqi1@huawei.com

Abstract

Multimodal Named Entity Recognition (MNER) uses supplementary images to resolve ambiguity in entity mentions in noisy, short-form social-media text. Prevailing approaches typically fuse text with visually aligned features, but such cues often lack sufficient discriminative power and may support conflicting entity-type hypotheses. To address this issue, we propose TCHV, a lightweight framework grounded in a type-conditioned hypothesis-verification perspective. By formulating each span-type pair as a distinct hypothesis during inference, TCHV predicts the latent feature evidence expected under each hypothesis and validates it against observed multimodal representations through bidirectional text-image matching, using a visual relevance gate to suppress uninformative images. Verification scores serve as surrogate evidence statistics; TCHV improves over established non-VLLM multimodal baselines on Twitter-2015/2017, and TCHV[†] leads the compared VLLM-based methods on Twitter-2017. Extensions demonstrate applicability to grounded and fine-grained settings. Code: <https://github.com/hanfeiyu703/ImaginE.git>.

1 Introduction

Multimodal Named Entity Recognition (MNER) aims to identify entity mentions in text-image social posts and assign them semantic types such as PER, LOC, ORG, and MISC (Moon et al., 2018; Lu et al., 2018; Zhang et al., 2018). Compared with text-only NER, MNER reflects a common social-media setting in which names, hashtags, abbreviations (Derczynski et al., 2015), and images jointly signal the entity being discussed (Moon et al., 2018; Lu et al., 2018; Adjali et al., 2020), supporting applications such as event monitoring (Adjali et al.,

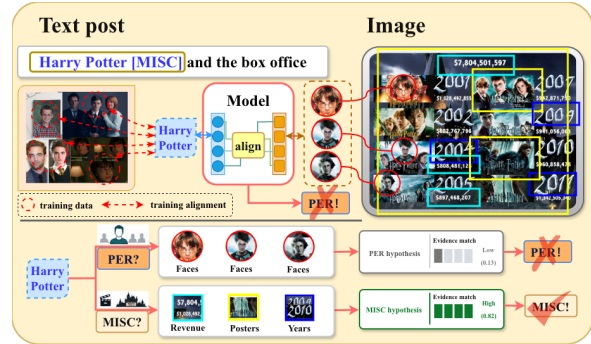


Figure 1: Motivation for TCHV. Visually related cues may support competing entity-type hypotheses. In the *Harry Potter* example, TCHV verifies candidate types through type-conditioned comparison of observed multimodal evidence. Scores are schematic.

2020), opinion analysis and information extraction (Derczynski et al., 2015), and multimodal search (Liu et al., 2023). Visual cues are useful because social-media texts are often short, noisy, and context-limited, making many mentions ambiguous from text alone (Moon et al., 2018; Lu et al., 2018; Chen et al., 2021).

Existing MNER methods usually follow a relevance-oriented *encode-align-fuse* paradigm: they first retrieve visually related evidence and then map the fused representation to an entity type (Xu et al., 2022; Wang et al., 2022; Chen et al., 2022; Zhou et al., 2022; Xu et al., 2024). However, type-discriminative decisions require distinguishing among candidate types that share visually related cues. Because cross-modal alignment is learned from training-time co-occurrence statistics and representation-level compatibility, it may over-emphasize visually salient or historically frequent cues that are related to a mention but insufficient for resolving its current entity type. This motivates type-conditioned evidence evaluation.

Figure 1 illustrates this distinction with the mention “*Harry Potter*” in the post “*Harry Potter and the box office*.” For example, in “*Harry Potter and the box office*,” character faces are clearly related to the Harry Potter universe and may induce a PER-like visual expectation. Yet the current context points to a movie-franchise interpretation, where movie posters, release years, revenue statistics, and the textual cue “*box office*” provide stronger evidence for MISC. This reveals a mismatch between relevance-oriented alignment and type-specific decision making: *the model must determine how observed multimodal evidence supports a particular candidate type in addition to measuring mention-image relatedness*.

To address this mismatch, we reformulate MNER as *type-conditioned hypothesis verification*. Instead of treating entity type only as a final classification label, we regard each span-type pair as a competing hypothesis. For each hypothesis, TCHV predicts the latent evidence expected under the given span and candidate type, and verifies this expectation against observed multimodal representations through bidirectional text-image matching. This turns visual information from a general relevance cue into type-specific evidence, enabling the model to compare competing hypotheses using label-aware evidence. Here, latent prediction denotes compact, type-conditioned evidence prediction for discriminative verification, following the latent-space perspective of prior predictive models (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024; Garrido et al., 2024).

Our contributions are summarized as follows:

- We propose TCHV, a lightweight type-conditioned hypothesis verification framework that evaluates each span-type hypothesis by comparing expected latent evidence with observed multimodal evidence through bidirectional verification and a visual relevance gate.
- We provide an analytical hypothesis-verification view that interprets verification scores as surrogate evidence statistics and explains how contrastive verification and bidirectional score separation support type-specific decisions.
- Experiments on Twitter-2015 and Twitter-2017 demonstrate consistent improvements

over established non-VLLM multimodal baselines. VLLM-enhanced and extension results further show compatibility with offline visual knowledge and applicability to grounded and fine-grained MNER.

2 Preliminaries

Span-based MNER. Let $\mathbf{x} = (x_1, \dots, x_n)$ denote a sentence paired with an image $\mathbf{I}$. Following span-based MNER formulations (Sohrab and Miwa, 2018; Zhou et al., 2022; Xu et al., 2024), we enumerate candidate spans $\mathcal{S}(\mathbf{x}) = \{u = (i, j) \mid 1 \leq i \leq j \leq n, j - i + 1 \leq L\}$ and assign each span a label $y \in \mathcal{Y} = \{y_0\} \cup \mathcal{Y}^+$, where y_0 is the non-entity class O and $\mathcal{Y}^+$ contains the entity types PER, LOC, ORG, and MISC. We write $K = |\mathcal{Y}^+|$ ($K = 4$ on both benchmarks) and $C = |\mathcal{Y}| = K + 1$, including O. Evaluation follows the strict span-level protocol: a prediction is correct only when both its boundary and non-O type match the gold annotation.

Latent verification intuition. Latent predictive models learn to predict target representations in latent space rather than reconstruct raw observations (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024). TCHV adapts this idea for MNER: for each span-type hypothesis, it predicts the latent evidence expected under that hypothesis and verifies the prediction against observed multimodal evidence. In our formulation, the predicted latent serves as a type-conditioned evidence expectation for hypothesis verification. Appendix B gives the detailed mapping from latent prediction to TCHV.

3 Our Method: TCHV

3.1 Overview

Inspired by statistical hypothesis testing, TCHV treats each candidate span-type pair as a hypothesis and uses type-conditioned verification scores as evidence statistics; Section 3.8 provides the formal interpretation and explanatory results.

TCHV operationalizes type-conditioned hypothesis verification through two latent prediction paths. The forward path evaluates whether the image contains evidence expected under a span-type hypothesis, while the reverse path evaluates whether visually grounded evidence remains consistent with the textual mention under the same type hypothesis. TCHV builds a type-conditioned verification layer on top of latent span and image representations:

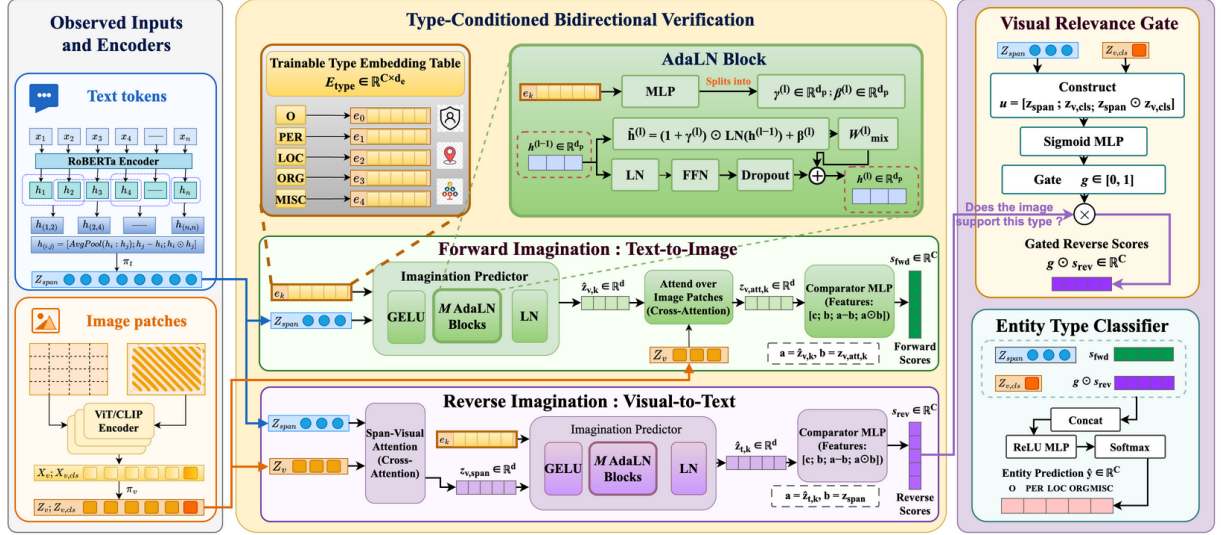


Figure 2: Architecture of TCHV. **(Left)** Backbone encoders produce span and visual representations. **(Center)** The forward path generates type-conditioned latent visual evidence expectations and compares them with observed image features. **(Right)** The reverse path predicts type-conditioned textual evidence expectations from span-relevant visual features and compares them with the real span, gated by visual relevance. The span subscript u is omitted in the diagram for readability.

cross-modal alignment retrieves related evidence, and verification evaluates its support for each candidate type. Figure 2 illustrates the full architecture. The pipeline has five stages: **(1)** Backbone encoders produce span representations $\mathbf{z}_{\text{span}} \in \mathbb{R}^d$, patch features $\mathbf{Z}_v \in \mathbb{R}^{N \times d}$, and a global visual vector $\mathbf{z}_{v,\text{cls}} \in \mathbb{R}^d$. **(2)** The forward path generates type-conditioned latent visual evidence expectations and scores each expectation against the observed image, yielding $\mathbf{s}_u^{\text{fwd}} \in \mathbb{R}^C$. **(3)** The reverse path predicts type-conditioned textual evidence expectations from visual features and scores each against the real span, yielding $\mathbf{s}_u^{\text{rev}} \in \mathbb{R}^C$. **(4)** A Visual Relevance Gate produces a scalar $g_u \in [0, 1]$ to down-weight reverse scores when the image is noisy, irrelevant, or weakly associated with the span. **(5)** An entity classifier fuses all signals to predict entity types.

3.2 Backbone Encoders

Text encoder. We encode the input sentence with RoBERTa (Liu et al., 2019), obtaining token-level representations $\mathbf{H}_t = (\mathbf{h}_1, \dots, \mathbf{h}_n) \in \mathbb{R}^{n \times d_t}$. For each candidate span (i, j) , we construct a span representation via the triple encoding scheme:

$$\mathbf{h}_{(i,j)} = [\text{AvgPool}(\mathbf{h}_i:\mathbf{h}_j) ; \mathbf{h}_j - \mathbf{h}_i ; \mathbf{h}_i \odot \mathbf{h}_j],$$

where $\text{AvgPool}(\mathbf{h}_i:\mathbf{h}_j) = \frac{1}{j-i+1} \sum_{\ell=i}^j \mathbf{h}_\ell$, $;$ denotes concatenation, and $\odot$ is the Hadamard product. The resulting $\mathbf{h}_{(i,j)} \in \mathbb{R}^{3d_t}$ is projected to the

shared latent space:

$$\mathbf{z}_{\text{span},u} = \pi_t(\mathbf{h}_{(i,j)}) \in \mathbb{R}^d, \quad \pi_t : \mathbb{R}^{3d_t} \rightarrow \mathbb{R}^d.$$

We write $\mathbf{z}_{\text{span}}$ when the span index u is unambiguous.

Image encoder. We encode image $\mathbf{I}$ with a Vision Transformer (Dosovitskiy et al., 2021), yielding patch representations $\mathbf{X}_v \in \mathbb{R}^{(N+1) \times d_v}$ including the [CLS] token. These are projected to the shared space:

$$\begin{aligned} \mathbf{Z}_v &= \pi_v(\mathbf{X}_{v,1:N}) \in \mathbb{R}^{N \times d}, \\ \mathbf{z}_{v,\text{cls}} &= \pi_v(\mathbf{x}_{v,0}) \in \mathbb{R}^d, \end{aligned}$$

where $\mathbf{X}_{v,1:N}$ denotes the N patch representations excluding the [CLS] token $\mathbf{x}_{v,0}$, and $\pi_v : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^d$ applies layer normalization followed by a linear projection. The RoBERTa and ViT backbone encoders are frozen in all experiments. We train the projection layers and the task-specific latent predictors, comparators, gate, and classifier modules.

3.3 Forward Latent Evidence Prediction: Text-to-Image Verification

The forward path is the main mechanism for type-evidence verification. Instead of directly attending from a span to visual patches in a type-agnostic manner, it first uses each type hypothesis to generate a latent visual evidence expectation and then verifies this expectation against the observed image.

It consists of a *Type-Conditioned Latent Evidence Predictor* that generates type-specific visual evidence expectations, and an *Evidence-Reality Comparator* that scores each expectation against the observed image.

Type-conditioned latent evidence predictor. We maintain a learnable label-hypothesis embedding table $\mathbf{E}_{\text{type}} \in \mathbb{R}^{C \times d_e}$, whose row $\mathbf{e}_k$ represents hypothesis $k \in \mathcal{Y}$; $\mathbf{e}_0$ encodes the O/null hypothesis. The null hypothesis supports non-entity rejection in classification, while semantic entity types in $\mathcal{Y}^+$ provide the evidence targets used by the auxiliary objectives. The predictor is a mapping $f_\theta: \mathbb{R}^d \times \mathbb{R}^{d_e} \rightarrow \mathbb{R}^d$. For each candidate span u and label hypothesis $k \in \mathcal{Y}$, it produces

$$\mathbf{a}_{u,k} \equiv \hat{\mathbf{z}}_{u,k}^v = f_\theta(\mathbf{z}_{\text{span},u}, \mathbf{e}_k).$$

Because $\mathbf{e}_k$ changes with k , the same span produces a distinct expected-evidence vector under each hypothesis.

We implement f_θ as an AdaLN-conditioned predictor with an input projection, M type-conditioned blocks, and an output projection. The type embedding modulates hidden states through adaptive affine parameters, allowing the same span representation to produce different latent evidence expectations under different type hypotheses. Appendix B gives the full block equations and initialization details.

Evidence-reality comparator. For each span-type hypothesis (u, k) , the comparator attends from the expected visual evidence $\mathbf{a}_{u,k}$ to the observed image patches $\mathbf{Z}_v$:

$$\begin{aligned} \alpha_{u,k} &= \text{softmax}\left(\frac{\mathbf{a}_{u,k} \mathbf{Z}_v^\top}{\sqrt{d}}\right) \in \mathbb{R}^N, \\ \mathbf{b}_{u,k} &\equiv \mathbf{z}_{u,k}^{v,\text{att}} = \alpha_{u,k} \mathbf{Z}_v \in \mathbb{R}^d. \end{aligned}$$

A scoring MLP produces a scalar compatibility score using common attend-and-compare matching features (Parikh et al., 2016; Chen et al., 2017).

$$\begin{aligned} \mathbf{m}_{u,k}^{\text{fwd}} &= [\mathbf{a}_{u,k}; \mathbf{b}_{u,k}; \mathbf{a}_{u,k} - \mathbf{b}_{u,k}; \mathbf{a}_{u,k} \odot \mathbf{b}_{u,k}], \\ s_{u,k}^{\text{fwd}} &= \text{MLP}_s(\mathbf{m}_{u,k}^{\text{fwd}}), \end{aligned} \quad (1)$$

where $\text{MLP}_s: \mathbb{R}^{4d} \rightarrow \mathbb{R}$. Thus, attention retrieves evidence under a type-conditioned expectation, while $s_{u,k}^{\text{fwd}}$ is the learned expected-versus-observed comparison rather than the raw attention similarity. The vector $\mathbf{s}_u^{\text{fwd}} = (s_{u,0}^{\text{fwd}}, \dots, s_{u,K}^{\text{fwd}}) \in \mathbb{R}^C$ contains one score per label hypothesis.

3.4 Reverse Consistency Verification: Visual-to-Text

The reverse path provides a complementary visual-to-text consistency verification signal. Given span-relevant visual evidence, it predicts the textual evidence expected under each type hypothesis and compares it with the actual span representation.

Span-visual attention. We extract a span-relevant visual representation via cross-attention with learned projections $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d \times d}$: The span representation serves as the query, while projected image patches serve as keys, so the attention distribution is conditioned on the current mention rather than on the image alone. This mention-level grounding is important for the reverse path because different spans in the same post may rely on different visual regions or may require suppressing the image entirely.

$$\begin{aligned} \mathbf{q}_u &= \mathbf{z}_{\text{span},u} \mathbf{W}_q, \quad \mathbf{K}_v = \mathbf{Z}_v \mathbf{W}_k, \\ \beta_u &= \text{softmax}\left(\frac{\mathbf{q}_u \mathbf{K}_v^\top}{\sqrt{d}}\right), \quad \mathbf{z}_{v,u} = \beta_u \mathbf{Z}_v. \end{aligned}$$

This lets each span attend to the image regions most relevant to it.

Reverse latent evidence predictor. A reverse predictor $r_\phi: \mathbb{R}^d \times \mathbb{R}^{d_e} \rightarrow \mathbb{R}^d$, architecturally identical to the forward predictor, takes $\mathbf{z}_{v,u}$ as input and generates an expected textual representation for each label hypothesis:

$$\hat{\mathbf{z}}_{u,k}^t = r_\phi(\mathbf{z}_{v,u}, \mathbf{e}_k), \quad k \in \mathcal{Y}.$$

To enforce consistent type semantics across both directions, the forward and reverse predictors *share the type embedding table* $\mathbf{E}_{\text{type}}$.

Reverse evidence comparator. The reverse comparator scores each $\hat{\mathbf{z}}_{u,k}^t$ against the real span $\mathbf{z}_{\text{span},u}$ using the same four-way feature combination as Eq. (1):

$$\begin{aligned} \mathbf{m}_{u,k}^{\text{rev}} &= [\hat{\mathbf{z}}_{u,k}^t; \mathbf{z}_{\text{span},u}; \hat{\mathbf{z}}_{u,k}^t - \mathbf{z}_{\text{span},u}; \\ &\quad \hat{\mathbf{z}}_{u,k}^t \odot \mathbf{z}_{\text{span},u}], \\ s_{u,k}^{\text{rev}} &= \text{MLP}_r(\mathbf{m}_{u,k}^{\text{rev}}), \end{aligned}$$

where $\text{MLP}_r: \mathbb{R}^{4d} \rightarrow \mathbb{R}$. The reverse scores $\mathbf{s}_u^{\text{rev}} \in \mathbb{R}^C$ reflect complementary type evidence from visual-to-text consistency verification.

The two branches are complementary. The forward branch evaluates whether the image contains evidence expected under a span-type hypothesis,

while the reverse branch evaluates whether visually grounded evidence remains consistent with the textual mention. The gate prevents the reverse branch from amplifying noise when the image is weakly related or irrelevant.

3.5 Visual Relevance Gate

Not every image is informative for entity typing; social media images may be irrelevant or even introduce visual bias/noise (Chen et al., 2021, 2022; Zhang et al., 2023). To prevent unreliable visual evidence from dominating the reverse path, we use a learned gate that estimates whether visual evidence is reliable enough to contribute to reverse verification. Let $\mathbf{u}_g = [\mathbf{z}_{\text{span},u}; \mathbf{z}_{v,\text{cls}}; \mathbf{z}_{\text{span},u} \odot \mathbf{z}_{v,\text{cls}}] \in \mathbb{R}^{3d}$:

$$g_u = \sigma(\text{MLP}_g(\mathbf{u}_g)) \in [0, 1], \quad (2)$$

where σ is the sigmoid function and $\text{MLP}_g: \mathbb{R}^{3d} \rightarrow \mathbb{R}$. The bias of the final layer is initialized to 2, so that $\sigma(2) \approx 0.88$: the gate defaults to “on” and learns to reduce the contribution of reverse scores when the image is irrelevant, noisy, or weakly associated with the span. The gated reverse scores are $\tilde{s}_{u,k}^{\text{rev}} = g_u s_{u,k}^{\text{rev}}$.

3.6 Entity Classifier

The classifier fuses five groups of features for span u : $\mathbf{z}_{\text{span},u}$, $\mathbf{z}_{v,\text{cls}}$, $\mathbf{s}_u^{\text{fwd}}$, $\tilde{\mathbf{s}}_u^{\text{rev}}$, and a span width embedding $\mathbf{w}_u \in \mathbb{R}^{d_w}$ indexed by span length. Let $\mathbf{c}_u = [\mathbf{z}_{\text{span},u}; \mathbf{z}_{v,\text{cls}}; \mathbf{s}_u^{\text{fwd}}; \tilde{\mathbf{s}}_u^{\text{rev}}; \mathbf{w}_u] \in \mathbb{R}^{2d+2C+d_w}$.

$$\begin{aligned} \mathbf{h}_{\text{fuse},u} &= \text{MLP}_{\text{fuse}}(\mathbf{c}_u) \in \mathbb{R}^{d_c}, \\ \mathbf{o}_u &= \mathbf{W}_{\text{cls}} \mathbf{h}_{\text{fuse},u} \in \mathbb{R}^C, \\ \hat{\mathbf{y}}_u &= \text{softmax}(\mathbf{o}_u) \in \mathbb{R}^C, \end{aligned}$$

where MLP_{fuse} is a two-layer network with GELU and dropout, $\mathbf{o}_u$ denotes pre-softmax logits, and $\hat{\mathbf{y}}_u$ denotes probabilities. At inference, we retain non-O spans whose best entity-type logit exceeds the O logit by a tuned margin, and greedily resolve overlapping predictions by confidence; Appendix F gives the decoding details.

3.7 Training Objective

The total loss combines a span classification loss, three forward verification losses, and their reverse counterparts:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{ira}} + \beta \mathcal{L}_{\text{ico}} + \gamma \mathcal{L}_{\text{sig}} \\ &\quad + \alpha' \mathcal{L}'_{\text{ira},g} + \beta' \mathcal{L}'_{\text{ico},g} + \gamma' \mathcal{L}'_{\text{sig}}, \end{aligned}$$

where $\mathcal{L}_{\text{task}}$ is weighted span-level cross-entropy; $\mathcal{L}_{\text{ira}}$, $\mathcal{L}_{\text{ico}}$, and $\mathcal{L}_{\text{sig}}$ denote evidence-reality alignment, type-contrastive verification, and anti-collapse regularization, respectively; primed symbols denote reverse counterparts. The subscript g indicates that reverse alignment and contrastive losses are weighted by the stopped-gradient visual relevance gate. Gold non-O entity spans provide the type-specific targets for auxiliary verification losses; batches without gold entity spans optimize $\mathcal{L}_{\text{task}}$ alone. The task loss supervises every enumerated candidate, including non-entities. At inference, the shared predictors and comparators process the full candidate set, and the classifier uses the learned entity-vs-O margin to identify entity spans. Appendix B gives the full objective definitions.

3.8 A Hypothesis-Verification View

Interpretation. For a candidate span u and a non-O type k , MNER can be viewed as testing whether the available multimodal evidence supports the span-type hypothesis against the composite null:

$$H_0^{u,k} : Y_u \neq k \quad \text{vs.} \quad H_1^{u,k} : Y_u = k.$$

TCHV constructs type-conditioned verification scores for these hypotheses. Let $F_{u,k} = s_{u,k}^{\text{fwd}}$ and $R_{u,k} = s_{u,k}^{\text{rev}}$. The forward score evaluates whether the observed image contains the visual evidence expected under (u, k) , while the reverse score evaluates whether span-relevant visual evidence remains consistent with the textual span under the same type hypothesis. With the visual relevance gate $g_u \in [0, 1]$, a reliability-weighted verification statistic can be abstracted as

$$T_{u,k} = F_{u,k} + \rho g_u R_{u,k}, \quad (3)$$

where $\rho \geq 0$ parameterizes the reverse term in the analytical statistic. The implemented classifier learns the corresponding fusion from the full feature vector through MLP_{fuse} ; Eq. (3) summarizes how forward verification, reverse consistency, and reliability gating jointly support type-specific decisions. This verification view characterizes the geometry of the learned scores: ρ serves as an analysis coefficient for the combined statistic, while operational predictions use the classifier logits and dev-selected entity-versus-O margin in Section 3.6. The propositions describe separation under their stated assumptions, and the empirical correspondence below examines the corresponding trends

across trained configurations and epochs. This perspective also clarifies why conditioning on the candidate type is necessary. Without type conditioning, visually related evidence for different labels is compared only after fusion, so evidence that is salient for one type can obscure weaker but more appropriate evidence for another. By constructing a separate statistic for each span-type hypothesis, TCHV makes the evidence comparison label-aware before the final classifier combines the signals.

Contrastive verification as a soft evidence-ratio surrogate. The forward contrastive loss encourages the gold type to dominate competing hypotheses. For a gold entity span u with type $y \in \mathcal{Y}^+$ and $C = |\mathcal{Y}|$, define

$$\bar{Z}_{u,y} = \frac{1}{C-1} \sum_{\ell \in \mathcal{Y} \setminus \{y\}} \exp\left(\frac{F_{u,\ell}}{\tau}\right),$$

$$\Lambda_{u,y}^{\text{fwd}} = \frac{F_{u,y}}{\tau} - \log \bar{Z}_{u,y}.$$

This quantity compares support for the correct type against a smooth aggregate of all competing labels, including the O/null class.

Proposition 1 (Contrastive loss as evidence-ratio margin). *For each gold entity span u with type $y \in \mathcal{Y}^+$, the instance-wise forward contrastive loss satisfies*

$$\begin{aligned} \ell_u^{\text{fwd}} &= -\log \frac{\exp(F_{u,y}/\tau)}{\sum_{k \in \mathcal{Y}} \exp(F_{u,k}/\tau)} \\ &= \log \left(1 + (C-1) \exp(-\Lambda_{u,y}^{\text{fwd}}) \right). \end{aligned} \quad (4)$$

Therefore, the instance-wise contrastive loss is a strictly decreasing function of $\Lambda_{u,y}^{\text{fwd}}$: smaller evidence-ratio margins are monotonically penalized, and minimizing the loss encourages larger support for the gold type relative to the composite null.

Proposition 1 shows that the contrastive objective monotonically rewards a larger soft evidence margin between the gold type and the composite set of competing hypotheses. Additional score-separation analysis is provided in Appendix C.

Bidirectional verification as combining two tests. Let $G_{u,k} = g_u R_{u,k}$ be the gated reverse statistic and define

$$\begin{aligned} \Delta_f &= \mathbb{E}[F_{u,k} \mid H_1^{u,k}] - \mathbb{E}[F_{u,k} \mid H_0^{u,k}], \\ \Delta_g &= \mathbb{E}[G_{u,k} \mid H_1^{u,k}] - \mathbb{E}[G_{u,k} \mid H_0^{u,k}]. \end{aligned}$$

Proposition 2 (Bidirectional verification can improve separation). *Assume $F_{u,k}$ and $G_{u,k} = g_u R_{u,k}$ are independent sub-Gaussian statistics under both hypotheses, with variance proxies σ_f^2 and σ_g^2 . If the forward statistic is discriminative, $\Delta_f > 0$, and the gated reverse statistic provides positive additional separation, $\Delta_g > 0$, then there exists a sufficiently small $\rho > 0$ such that*

$$\frac{(\Delta_f + \rho \Delta_g)^2}{\sigma_f^2 + \rho^2 \sigma_g^2} > \frac{\Delta_f^2}{\sigma_f^2}.$$

Proposition 2 gives a testing-theoretic explanation of bidirectional complementarity: reverse verification is useful when it supplies additional separation beyond the forward score, but it can add variance when the image is unreliable. The visual relevance gate therefore acts as a reliability controller, allowing reverse evidence to contribute when informative and suppressing it when visual evidence is weak or noisy.

Empirical correspondence. Across seven trained configurations, two data splits, and four entity types ($N = 56$), the mean gold-type forward separation $F_{u,y} - \max_{j \in \mathcal{Y}^+ \setminus \{y\}} F_{u,j}$ correlates with forward-score argmax accuracy (Pearson $r = 0.730$). Across 840 epoch-level records, the forward InfoNCE training proxy $I_{\text{NCE}}^{\text{proxy}} = \log K - \mathcal{L}_{\text{ico}}$, with $K = 4$ non-O types, correlates with epoch-level F1 ($r = 0.904$). The two correlations support the link between score separation, contrastive optimization, and entity-type discrimination.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate standard MNER on Twitter-2015 and Twitter-2017, two multimodal social media benchmarks annotated with PER, LOC, ORG, and MISC entities (Lu et al., 2018; Zhang et al., 2018). Following standard protocol, we report Precision (P), Recall (R), F1, and per-type F1 under strict span-level matching. Table 2 summarizes the benchmark splits and annotation counts; extension benchmark protocols are provided in Appendix F.

Implementation details. The default TCHV uses frozen RoBERTa-base (Liu et al., 2019) and ViT-Base/16 (Dosovitskiy et al., 2021) encoders, and trains only the task-specific projection, latent

Table 1: Performance on Twitter-2015 and Twitter-2017. Per-type F1 and overall P, R, F1 (%). Best and second-best results are bolded and underlined separately within the multimodal MNER and VLLM-based groups. The dagger denotes VLLM-enhanced results. Within the VLLM group, PGIM leads on Twitter-2015 and TCHV[†] leads on Twitter-2017.

Method	Twitter-2015							Twitter-2017						
	PER	LOC	ORG	MISC	P	R	F1	PER	LOC	ORG	MISC	P	R	F1
<i>Text-only methods</i>														
BiLSTM-CRF	76.77	72.56	41.33	26.80	68.14	61.09	64.42	85.12	72.68	72.50	52.56	79.42	73.43	76.31
BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
<i>Multimodal MNER methods</i>														
UMT	85.24	81.58	63.03	39.45	71.67	75.23	73.41	91.56	84.73	82.24	70.10	85.28	85.34	85.31
MAF	84.67	81.18	63.35	41.82	71.86	75.10	73.42	91.51	85.80	85.10	68.79	86.13	86.38	86.25
HVPNeT	86.62	81.89	60.22	41.70	73.87	76.82	75.32	92.87	84.75	83.24	69.35	85.84	87.93	86.87
MNER-QG	85.31	81.65	63.41	41.32	77.43	72.15	74.70	92.92	86.19	84.52	71.67	88.26	85.65	86.94
De-Bias	85.97	81.84	64.02	43.38	74.45	76.13	75.28	93.46	84.15	84.42	67.88	87.59	86.11	86.84
SMNER	85.98	81.26	64.53	43.26	75.34	76.81	76.06	92.63	85.62	85.68	71.71	88.15	87.47	87.81
VEC-MNER	86.11	81.03	62.86	40.60	74.56	75.23	74.89	93.88	81.27	85.49	73.40	87.42	87.61	87.51
ICKA	87.01	83.85	65.87	48.28	72.36	78.75	75.42	93.99	87.24	86.24	<u>75.76</u>	85.13	89.19	87.12
SEPA	86.49	83.33	62.90	43.80	74.71	76.84	75.76	93.18	86.27	85.39	<u>69.08</u>	87.50	87.05	87.27
RPI-HMIF	86.90	82.48	64.82	45.83	77.54	76.02	76.77	93.71	86.04	<u>87.45</u>	71.48	89.45	87.27	88.35
TCHV-BERT (Ours)	<u>87.53</u>	83.15	<u>67.34</u>	<u>48.50</u>	<u>77.72</u>	<u>77.33</u>	<u>77.53</u>	95.31	84.57	86.99	73.38	<u>89.84</u>	88.54	<u>89.19</u>
TCHV (Ours)	87.71	<u>83.57</u>	68.16	50.25	78.35	<u>78.09</u>	78.22	<u>94.85</u>	<u>87.03</u>	88.18	79.44	90.64	89.76	90.20
<i>VLLM-based methods</i>														
Llama3.2-11B-Vision (zero-shot)	75.43	68.98	36.52	26.05	49.29	70.72	58.09	85.48	63.37	61.05	24.36	58.78	71.77	64.63
Llama3.2-11B-Vision (4-shot)	76.98	68.23	37.57	24.15	50.35	70.20	58.64	87.28	65.91	66.41	30.15	64.19	74.83	69.10
Qwen2.5-VL-7B (zero-shot)	67.57	58.68	34.20	18.25	51.76	51.81	51.78	77.72	53.87	51.90	18.60	63.31	55.79	59.31
Qwen2.5-VL-7B (4-shot)	67.83	59.26	36.89	20.78	55.42	58.30	56.82	79.08	55.71	53.67	19.14	65.48	59.75	62.48
DeepSeek-VL-7B (zero-shot)	52.74	27.92	16.37	6.71	21.44	27.87	24.23	63.14	29.29	15.79	11.91	33.21	33.61	33.41
DeepSeek-VL-7B (4-shot)	57.04	28.01	18.83	9.13	26.48	31.51	28.78	64.27	47.92	38.01	12.99	35.33	49.89	41.36
PGIM	88.34	84.22	<u>70.15</u>	<u>52.34</u>	79.21	79.45	79.33	<u>96.46</u>	89.89	<u>89.03</u>	<u>79.62</u>	<u>90.86</u>	<u>92.01</u>	<u>91.43</u>
TCHV [†] (Ours)	<u>88.16</u>	<u>83.39</u>	70.42	52.75	<u>78.50</u>	<u>79.29</u>	<u>78.89</u>	96.88	<u>88.63</u>	90.00	79.86	91.02	92.90	91.95

Table 2: Dataset statistics. Train/Dev/Test report text-image posts; Ent. and Grd. report total entity and groundable-entity annotations when available.

Dataset	Train	Dev	Test	Ent.	Grd.
Twitter-2015	4,000	1,000	3,257	–	–
Twitter-2017	3,373	723	723	–	–
Twitter-GMNER	7,000	1,500	1,500	16,778	6,716
Twitter-FMNERG	7,000	1,500	1,500	16,772	6,770

verification, gate, and classifier modules. TCHV-BERT replaces only RoBERTa-base with BERT-base (Devlin et al., 2019); otherwise the visual encoder, trainable modules, training protocol, and decoding settings are unchanged. Full hyperparameters, preprocessing, decoding, and computing environment details are provided in Appendix F.

Baselines. We compare against three groups of methods: (1) *Text-only*: BiLSTM-CRF (Huang et al., 2015) and BERT-CRF (Devlin et al., 2019), with the Twitter-2015/2017 results in Table 1 taken from Yu et al. (Yu et al., 2020); (2) *Multimodal MNER*: UMT (Yu et al., 2020), MAF (Xu et al., 2022), HVPNeT (Chen et al., 2022), MNER-QG (Jia et al., 2023), De-Bias (Zhang et al., 2023), SMNER (Xu et al., 2024), VEC-MNER (Wei et al., 2024), ICKA (Zeng et al., 2024), SEPA (Ding et al., 2025), and RPI-HMIF (Wei et al., 2026); (3) *VLLM-*

based: Llama3.2-11B-Vision (Meta AI, 2024), Qwen2.5-VL-7B (Bai et al., 2025), DeepSeek-VL-7B (Lu et al., 2024), and PGIM (Li et al., 2023a). The ICKA and RPI-HMIF entries in Table 1 reproduce the results reported in their original papers, so Table 1 provides a benchmark-level comparison across published configurations. Table 4 complements it with shared frozen backbones and an identical training and decoding pipeline for the models compared there.

4.2 Main Results

Table 1 presents the main comparison, organized by methodology.

The results show that type-conditioned verification provides evidence beyond relevance-oriented alignment: TCHV improves overall F1 over text-only and multimodal baselines on both datasets. Among VLLM-enhanced methods, PGIM leads on Twitter-2015 (79.33 vs. 78.89 F1), while TCHV[†] leads on Twitter-2017 (91.95 vs. 91.43). Per-type gains are strongest on MISC, where visual relatedness often fails to determine a single canonical entity type. This pattern is consistent with our motivation: MISC contains events, products, competitions, and media titles, so visually related cues can easily point to people, organizations, or

 Pat celebrating her 90th birthday with Emily Roux in Chez Roux at the Newmarket Guineas Festival. #weare350 (1) Pat:PER (2) Emily Roux:PER (3) Chez Roux:LOC (4) Newmarket Guineas Festival:MISC MAF: (1) ✓ (2) ✓ (3) ✗ Roux:MISC (4) ✓ UMT: (1) ✓ (2) ✓ (3) ✗ Chez Roux:MISC (4) ✓ TCHV: (1) ✓ (2) ✓ (3) ✓ (4) ✓	 David Beckham sends message to Zinedine Zidane following Real Madrid's UCL win. Class! (1) David Beckham:PER (2) Zinedine Zidane:PER (3) Real Madrid:ORG (4) UCL:MISC MAF: (1) ✓ (2) ✓ (3) ✓ (4) ✗ UCL:ORG UMT: (1) ✓ (2) ✓ (3) ✓ (4) ✗ UCL:ORG TCHV: (1) ✓ (2) ✓ (3) ✓ (4) ✓
--	--

Figure 3: Case study. TCHV resolves type ambiguity by verifying candidate entity-type hypotheses.

Table 3: Ablation study on Twitter-2017 (F1, %). Δ is the change from the full TCHV model.

Variant	F1	Δ
Full TCHV	90.20	–
<i>Architecture ablations</i>		
w/o Forward Prediction	89.14	-1.06
w/o Reverse Verification	89.41	-0.79
w/o Visual Relevance Gate	89.07	-1.13
w/o AdaLN (type concat.)	88.98	-1.22
Replace Latent Pred. with Cross-Attn	89.67	-0.53
<i>Loss ablations</i>		
w/o $\mathcal{L}_{ira}$ (alignment)	89.46	-0.74
w/o $\mathcal{L}_{ico}$ (contrastive)	89.10	-1.10
w/o $\mathcal{L}_{sig}$ (anti-collapse)	89.30	-0.90
w/o Stop-Gradient in $\mathcal{L}_{ira}$	89.91	-0.29

locations unless they are checked under the candidate type. The TCHV-BERT variant reaches 77.53/89.19 F1, showing that the verification framework is effective with both BERT and RoBERTa text encoders. Meanwhile, direct zero-shot or few-shot VLLM prompting remains weaker than task-specific MNER models, indicating that large vision-language models still benefit from span-level supervision and explicit type verification in this setting.

4.3 Ablation Study

Table 3 reports ablation results on Twitter-2017. Ablations assess both verification paths, the gate, type conditioning, and auxiliary objectives.

Removing reverse verification lowers F1 by 0.79 points, showing that the gated reverse path contributes complementary consistency evidence. Retaining the reverse branch while removing the visual relevance gate produces a 1.13-point drop, supporting reliability-weighted integration of this evidence. Appendix F.4 further evaluates the gate under four initialization priors. Together with the type-conditioning and objective ablations, these results support the complete verification design.

4.4 Type Discrimination and Same-Backbone Controls

We quantify type discrimination in the forward score by predicting the type of each gold non-O test span with $\arg \max_{k \in \mathcal{Y}^+} s_{u,k}^{\text{fwd}}$. Learned type embeddings reach 87.76% accuracy on Twitter-2015 and 94.84% on Twitter-2017. Mapping every type to a different learned embedding with a fixed cyclic permutation reduces accuracy to 4.54% and 2.09%, supporting a learned correspondence between candidate types and expected evidence. This forward-only diagnostic conditions on gold boundaries and isolates the type–embedding correspondence within a fixed checkpoint, complementing the end-to-end MNER F1 results.

The same-backbone comparison holds the frozen encoders and the training/decoding pipeline fixed. Relative to direct fusion, concatenation, cross-attention, and AdaLN improve F1 by 2.43/2.59, 2.69/2.71, and 3.02/3.24 points on Twitter-2015/2017, respectively. The consistent gains across all three conditioning operators support the shared type-conditioned verification interface, while AdaLN provides the strongest tested parameterization.

Mechanistic synthesis. Taken together, Tables 3 and 4 provide a staged mechanism analysis. The cyclic diagnostic links candidate-type identity to forward verification scores; the shared-backbone comparison shows gains for three conditioning operators; and the ablations associate the reverse path’s contribution with reliability gating. The correlations in Section 3.8 further connect stronger forward separation with type selection and the contrastive training proxy with epoch-level F1. This evidence supports type-conditioned score construction, operator-robust verification, and reliability-weighted bidirectional evidence as three empirically grounded roles within TCHV.

4.5 Extension to GMNER and FMNERG

TCHV extends to grounded and fine-grained settings by replacing patch-level visual evidence with proposal-level evidence and adding a lightweight region pointer, while retaining the span encoder and type-hypothesis scoring interface. Full protocols and P/R/F1 breakdowns are in Appendix E.

4.6 Case Study

Figure 3 shows two failure modes of relevance-oriented fusion. The LOC hypothesis separates



Figure 4: Type-evidence visualization. For the same image and mention, different entity-type hypotheses attend to different visual evidence, illustrating how type-conditioned verification separates plausible but competing type interpretations.

(a) Forward type diagnosis			(b) Same-backbone comparison				
Type embeddings	T15 Acc.	T17 Acc.	Model	Conditioning	HV	T15	T17
Learned	87.76 (4547/5181)	94.84 (1269/1338)	Direct fusion	None	No	75.20	86.96
Cyclic permutation	4.54 (235/5181)	2.09 (28/1338)	TCHV-Concat	Concatenation	Yes	77.63	89.55
			TCHV-Cross-attn.	Cross-attention	Yes	77.89	89.67
			TCHV-AdaLN	AdaLN	Yes	78.22	90.20

Table 4: Controlled analyses on Twitter-2015/2017. (a) Forward-score type accuracy on gold non-O spans with known boundaries. (b) F1 under the same frozen backbones and training/decoding pipeline; HV denotes hypothesis verification. Panel (b) is a separate controlled suite, while Table 3 retains the original component-ablation runs.

Table 5: Compact extension results. We report main-task F1 (%); full results are in Appendix E.

Method	GMNER	FMNERG
H-Index (Yu et al., 2023)	56.41	46.55
TIGER (Wang et al., 2023)	–	47.20
TCHV + Extension	59.73	50.64

Table 6: Parameter count and per-epoch runtime of TCHV. Trainable parameters exclude frozen backbone encoders.

Parameter count		
Trainable task-specific params	~15M	
Setting	Train only (s/epoch)	Train+Eval (s/epoch)
Twitter-2015	35.8	51.8
Twitter-2017	30.5	43.6
GMNER	136.1	152.8
FMNERG	431.2	469.0

“Chez Roux” from nearby person mentions, while verification assigns the event-like “UCL” mention to MISC despite organization-related football imagery. In both cases, the candidate-type hypothesis determines how related visual context contributes to the label.

Type-Evidence Visualization. Figure 4 shows that red-carpet context is visually related to people, but the model still verifies whether the event-oriented MISC hypothesis is better supported. The visualization supports the qualitative claim that type conditioning redistributes visual evidence across competing candidate hypotheses. The accompanying score profile makes the resulting preference explicit across all four candidate types.

4.7 Efficiency Analysis

Table 6 shows that TCHV trains about ~15M task-specific parameters while freezing both encoders; its added cost mainly comes from scoring the small type-hypothesis set. Longer GMNER/FMNERG runtimes track larger inputs and region-level evaluation, while the shared verification modules remain unchanged.

5 Conclusion

This paper proposes TCHV, a lightweight type-conditioned hypothesis-verification framework that evaluates each span–type hypothesis through bidirectional expected-versus-observed evidence comparison with a visual relevance gate. Experiments on Twitter-2015 and Twitter-2017 show that TCHV consistently improves over established non-VLLM multimodal baselines. Among the compared VLLM-based methods, PGIM leads on Twitter-2015 and TCHV[†] leads on Twitter-2017. Further extensions demonstrate its applicability to grounded and fine-grained settings. Across the grounded and fine-grained extensions, TCHV retains its span–type scoring interface while adapting visual evidence to proposal-level outputs. This shared formulation supports richer output structures while preserving a common verification objective across evidence granularities. The type diagnostics and same-backbone controls further show that explicit type-conditioned verification produces discriminative scores and improves over direct fusion under shared backbones.

6 Limitations

TCHV still depends on frozen textual and visual encoders, so its verification quality is bounded by the representations provided by these pretrained backbones. Extremely noisy, irrelevant, misleading, or underspecified images may still provide unreliable evidence, especially when the decisive cue is a small region, embedded text, or a subtle visual relation that is not well captured by global image features. The current design mainly verifies latent global or span-level evidence; finer region-level verification, explicit OCR-aware reasoning, and end-to-end grounding remain underexplored. Our extensions to grounded and fine-grained settings also depend on external proposal quality and predefined label inventories. More broadly, TCHV assumes a relatively small set of candidate entity types can be enumerated during inference, and scaling the same verification idea to open or very large type ontologies may require candidate retrieval, pruning, or calibration mechanisms. Generalization to other languages, platforms, domains, and distribution shifts therefore requires further study. Finally, the current hypothesis-verification view explains surrogate verification scores; future work can add calibrated uncertainty estimation and study reliability under stronger out-of-domain shifts.

7 Ethical Considerations

This work uses existing public multimodal NER benchmarks and does not collect new data or involve human participants. Since these benchmarks are derived from social-media posts, they may contain user-generated text, named entities, and images of real individuals, which raises privacy and consent concerns. We therefore intend our model and released code for research use only, and users should follow the licenses and terms of the original datasets.

TCHV may inherit biases from pretrained encoders and benchmark data, and its predictions may be unreliable for underrepresented groups, languages, platforms, or visually ambiguous cases. Misuse in surveillance, profiling, content moderation, or other high-stakes settings could cause harm. We discourage such deployment without additional privacy review, bias evaluation, and human oversight. To reduce computational cost, TCHV freezes large backbone encoders and trains only lightweight task-specific modules, but reproducing the experiments still requires GPU resources.

Acknowledgements

This work was partially supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB0680101) and the National Natural Science Foundation of China (No. 62472416). The model training was performed on the robotic AI-Scientist platform of Chinese Academy of Science.

References

- Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. [Building a multimodal entity linking dataset from tweets](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4285–4292, Marseille, France. European Language Resources Association.
- Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. [Self-supervised learning from images with a joint-embedding predictive architecture](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. [Layer normalization](#). *arXiv preprint arXiv:1607.06450*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. [Qwen2.5-VL technical report](#). *arXiv preprint arXiv:2502.13923*.
- Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. 2024. [Revisiting feature prediction for learning visual representations from video](#). *arXiv preprint arXiv:2404.08471*.
- Adrien Bardes, Jean Ponce, and Yann LeCun. 2022. [VICReg: Variance-invariance-covariance regularization for self-supervised learning](#). In *International Conference on Learning Representations*.
- Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. [Enhanced LSTM for natural language inference](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.
- Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. [Can images help recognize entities? a study of the role of images for multimodal](#)

- NER. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 87–96, Online. Association for Computational Linguistics.
- Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. [Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618, Seattle, United States. Association for Computational Linguistics.
- Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. [Analysis of named entity recognition and linking for tweets](#). *Information Processing & Management*, 51(2):32–49.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Guohui Ding, Yushuo Kong, and Xinlei Li. 2025. [Sepa: An semantic projection alignment framework for multimodal named entity recognition](#). In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 202–210. Association for Computing Machinery.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *International Conference on Learning Representations*.
- Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. 2024. [Learning and leveraging world models in visual representation learning](#). *arXiv preprint arXiv:2403.00504*.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. [Bidirectional lstm-crf models for sequence tagging](#). *arXiv preprint arXiv:1508.01991*.
- Meihuizi Jia, Lei Shen, Xin Shen, Lejian Liao, Meng Chen, Xiaodong He, Zhendong Chen, and Jiaqi Li. 2023. [MNER-QG: An end-to-end MRC framework for multimodal named entity recognition with query grounding](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8032–8040.
- Yann LeCun. 2022. [A path towards autonomous machine intelligence](#). *OpenReview preprint*.
- Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. 2023a. [Prompting chatgpt in mner: Enhanced multimodal named entity recognition with auxiliary refined knowledge](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2787–2802.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhui Chen, and William Wang. 2023. [EDIS: Entity-driven image search over multimodal web content](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4877–4894, Singapore. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. [Visual attention model for name tagging in multimodal social media](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, Melbourne, Australia. Association for Computational Linguistics.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. [Deepseek-vl: towards real-world vision-language understanding](#). *arXiv preprint arXiv:2403.05525*.
- Meta AI. 2024. [Llama 3.2 Vision Model Card](#).
- Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. [Multimodal named entity recognition for short social media posts](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860, New Orleans, Louisiana. Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *arXiv preprint arXiv:1807.03748*.
- Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. [A decomposable attention model for natural language inference](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

- William Peebles and Saining Xie. 2023. [Scalable diffusion models with transformers](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. [FiLM: Visual reasoning with a general conditioning layer](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Mohammad Golam Sohrab and Makoto Miwa. 2018. [Deep exhaustive model for nested named entity recognition](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849, Brussels, Belgium. Association for Computational Linguistics.
- Jieming Wang, Ziyang Li, Jianfei Yu, Li Yang, and Rui Xia. 2023. [Fine-grained multimodal named entity recognition and grounding with a generative framework](#). In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pages 3934–3943, New York, NY, USA. Association for Computing Machinery.
- Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022. [ITA: Image-text alignments for multi-modal named entity recognition](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189, Seattle, United States. Association for Computational Linguistics.
- Ziqi Wang, Chen Zhu, Zhi Zheng, Xinhang Li, Tong Xu, Yongyi He, Qi Liu, Ying Yu, and Enhong Chen. 2024. [Granular entity mapper: Advancing fine-grained multimodal named entity recognition and grounding](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3211–3226, Miami, Florida, USA. Association for Computational Linguistics.
- Pengfei Wei, Hongjun Ouyang, Qintai Hu, Bi Zeng, Guang Feng, and Qingpeng Wen. 2024. [VEC-MNER: Hybrid transformer with visual-enhanced cross-modal multi-level interaction for multimodal NER](#). In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 469–477. Association for Computing Machinery.
- Pengfei Wei, Binghan Zhang, Qingpeng Wen, Qintai Hu, Bi Zeng, and Guang Feng. 2026. [How people read? reading preference-inspired multimodal NER with heterogeneous mining and iterative fusion engine](#). *Information Fusion*, 127:103824.
- Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. [MAF: A general matching and alignment framework for multimodal named entity recognition](#). In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pages 1215–1223, New York, NY, USA. Association for Computing Machinery.
- Yongxiu Xu, Hao Xu, Heyan Huang, Shiyao Cui, Minghao Tang, Longzheng Wang, and Hongbo Xu. 2024. [An effective span-based multimodal named entity recognition with consistent cross-modal alignment](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1063–1072, Torino, Italia. ELRA and ICCL.
- Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. [Improving multimodal named entity recognition via entity span detection with unified multimodal transformer](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352, Online. Association for Computational Linguistics.
- Jianfei Yu, Ziyang Li, Jieming Wang, and Rui Xia. 2023. [Grounded multimodal named entity recognition on social media](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9141–9154, Toronto, Canada. Association for Computational Linguistics.
- Qingyang Zeng, Minghui Yuan, Jing Wan, Kunfeng Wang, Nannan Shi, Qianzi Che, and Bin Liu. 2024. [ICKA: An instruction construction and knowledge alignment framework for multimodal named entity recognition](#). *Expert Systems with Applications*, 255:124867.
- Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. [VinVL: Revisiting visual representations in vision-language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.
- Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. [Adaptive co-attention network for named entity recognition in tweets](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 2023. [Reducing the bias of visual objects in multimodal named entity recognition](#). In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, New York, NY, USA. Association for Computing Machinery.
- Baohang Zhou, Ying Zhang, Kehui Song, Wenya Guo, Guoqing Zhao, Hongbin Wang, and Xiaojie Yuan. 2022. [A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition](#). In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 6293–6302.

Appendix

A Related Work

Multimodal named entity recognition. Early MNER work studied multimodal social-media entity recognition and attention-based visual integration (Moon et al., 2018; Lu et al., 2018; Zhang et al., 2018). Later methods improved cross-modal interaction through unified multimodal transformers (Yu et al., 2020), matching and alignment frameworks (Xu et al., 2022; Wang et al., 2022), visual-context analysis and hierarchical visual prefixes (Chen et al., 2021, 2022), machine reading comprehension with query grounding (Jia et al., 2023), and debiasing strategies (Zhang et al., 2023). Recent span-level approaches include semi-supervised span-based MNER (Zhou et al., 2022), cross-modal span alignment (Xu et al., 2024), and semantic projection (Ding et al., 2025). Recent methods further incorporate task guidance into multimodal fusion: ICKA constructs multimodal instructions with VLM-derived external knowledge, whereas RPI-HMIF uses explicit questions for heterogeneous feature mining and iterative routing-based fusion (Zeng et al., 2024; Wei et al., 2026). These approaches focus on instruction and interaction design; TCHV is complementary in conditioning verification on each candidate entity type. Prior MNER studies improve visual-textual interaction through cross-modal attention, alignment, fusion, prompt-based conditioning, and noise-aware visual filtering. TCHV complements these relevance-oriented interactions by evaluating retrieved evidence under each candidate entity type, turning visual information into a type-conditioned verification signal.

Latent prediction and JEPA. JEPA-style models predict target latent representations rather than reconstructing pixels (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024). Recent work further studies latent world models and anti-collapse regularization (Garrido et al., 2024; Bardes et al., 2022). Our type-conditioned predictor uses adaptive conditioning related to FiLM and adaptive layer normalization (Perez et al., 2018; Peebles and Xie, 2023). Our work uses latent prediction in a task-specific sense: it predicts expected evidence under an entity-type hypothesis rather than generating images or modeling general environment dynamics.

Table 7: Mapping from JEPA-style latent prediction to entity-type verification in TCHV.

JEPA-style Prediction	TCHV (Ours)
Context repr. $\mathbf{z}_{\text{ctx}}$	Span representation
Conditioning $\mathbf{c}$	Entity-type or null hypothesis
Predicted target $\hat{\mathbf{z}}_{\text{tar}}$	Expected visual evidence
Encoded target $\mathbf{z}_{\text{tar}}$	Observed visual evidence
Prediction loss	Evidence-reality alignment $\mathcal{L}_{\text{ira}}$
Anti-collapse reg.	Variance/covariance regularizer $\mathcal{L}_{\text{sig}}$

Table 8: Core notation used in the verification modules and analysis.

Symbol	Meaning
$\mathcal{Y}^+, K$	non-O entity types and their count, $K = \mathcal{Y}^+ $
$\mathcal{Y}, C$	full label set and size, $\mathcal{Y} = \{y_0\} \cup \mathcal{Y}^+$ and $C = K + 1$
u, k	candidate span and label hypothesis; $k = 0$ denotes O
$\mathbf{z}_{\text{span}, u}$	representation of candidate span u
$\mathbf{a}_{u, k}$	expected visual evidence under hypothesis (u, k)
$\mathbf{b}_{u, k}$	attended observed evidence under hypothesis (u, k)
$s_{u, k}^{\text{fwd}}$	forward expected-observed evidence score
$s_{u, k}^{\text{rev}}$	reverse consistency score
$\tilde{s}_{u, k}^{\text{rev}}$	gated reverse score, $g_u s_{u, k}^{\text{rev}}$
g_u	span-level visual reliability gate
ρ	analytical weight of $g_u R_{u, k}$ in $T_{u, k}$
$\mathbf{w}_u$	embedding of the width of span u

B Additional Method Details

Table 7 summarizes how TCHV adapts latent predictive modeling to type-conditioned entity verification.

B.1 Predictor Architecture

The type-conditioned latent evidence predictor consists of an input projection, M AdaLN blocks, and an output projection. The input projection maps span u into an intermediate space shared across its label hypotheses:

$$\mathbf{h}_u^{(0)} = \sigma(\mathbf{W}_{\text{in}} \mathbf{z}_{\text{span}, u} + \mathbf{b}_{\text{in}}),$$

$$\mathbf{h}_{u, k}^{(0)} = \mathbf{h}_u^{(0)} \in \mathbb{R}^{d_p}.$$

where σ is GELU. Each AdaLN block $\ell \in \{1, \dots, M\}$ performs:

$$(\gamma_{\ell, k}, \beta_{\ell, k}) = \text{MLP}_{\ell}(\mathbf{e}_k), \quad (5)$$

$$\begin{aligned} \bar{\mathbf{h}}_{u, k}^{(\ell)} &= (1 + \gamma_{\ell, k}) \odot \text{LN}(\mathbf{h}_{u, k}^{(\ell-1)}) + \beta_{\ell, k}, \\ \tilde{\mathbf{h}}_{u, k}^{(\ell)} &= \mathbf{h}_{u, k}^{(\ell-1)} + \mathbf{W}_{\ell}^{\text{mix}} \bar{\mathbf{h}}_{u, k}^{(\ell)}, \\ \mathbf{h}_{u, k}^{(\ell)} &= \tilde{\mathbf{h}}_{u, k}^{(\ell)} + \text{FFN}(\text{LN}(\tilde{\mathbf{h}}_{u, k}^{(\ell)})), \end{aligned} \quad (6)$$

where LN is Layer Normalization (Ba et al., 2016); the type-conditioned affine modulation is related to feature-wise conditioning and adaptive layer normalization (Perez et al., 2018; Peebles and Xie, 2023). Here, $\gamma_\ell, \beta_\ell \in \mathbb{R}^{d_p}$ are affine parameters generated from type hypothesis $\mathbf{e}_k$, $\mathbf{W}_\ell^{\text{mix}} \in \mathbb{R}^{d_p \times d_p}$ is a linear mixing layer, and FFN is a two-layer network with GELU and expansion factor 2. The output projection yields $\hat{\mathbf{z}}_{u,k}^v = \mathbf{W}_{\text{out}} \text{LN}(\mathbf{h}_{u,k}^{(M)}) \in \mathbb{R}^d$.

The last linear layer of the conditioning MLP in Eq. (5) is zero-initialized, following adaptive conditioning practice in Diffusion Transformers (Peebles and Xie, 2023), so that type conditioning takes effect progressively. Since each predictor input is a single vector, self-attention provides no token-mixing benefit; we replace it with the linear mixing layer in Eq. (6).

B.2 Training Objective Details

Let $\mathcal{Y} = \{y_0\} \cup \mathcal{Y}^+$, where y_0 is O, $K = |\mathcal{Y}^+|$, and $C = |\mathcal{Y}| = K + 1$. Let $\mathcal{E} = \{s = (i, j) \in \mathcal{S}(\mathbf{x}) : y_s \in \mathcal{Y}^+\}$ be the set of gold entity spans. The task classification loss is computed over all candidate spans and the full label set:

$$\mathcal{L}_{\text{task}} = -\frac{1}{|\mathcal{S}|} \sum_{(i,j) \in \mathcal{S}} \sum_{k=0}^{C-1} w_k \tilde{y}_k^{(i,j)} \log \hat{y}_k^{(i,j)},$$

where $\tilde{y}_k^{(i,j)}$ is the label-smoothed target and $w_0 < 1$ down-weights the dominant non-entity class.

For non-O entity spans, the evidence-reality alignment loss aligns the predicted visual evidence expectation of the ground-truth entity type with the attended real visual evidence:

$$\mathcal{L}_{\text{ira}} = \frac{1}{|\mathcal{E}|} \sum_{u=(i,j) \in \mathcal{E}} \|\hat{\mathbf{z}}_{u,y_u}^v - \text{sg}(\mathbf{z}_{u,y_u}^{\text{att}})\|_2^2, \quad (7)$$

where $\text{sg}(\cdot)$ stops gradient flow through its argument. The evidence contrastive loss is an InfoNCE-style objective (Oord et al., 2018):

$$\mathcal{L}_{\text{ico}} = -\frac{1}{|\mathcal{E}|} \sum_{u=(i,j) \in \mathcal{E}} \log \frac{\exp(s_{u,y_u}^{\text{fwd}}/\tau)}{\sum_{k=0}^{C-1} \exp(s_{u,k}^{\text{fwd}}/\tau)},$$

where $\tau > 0$ is a temperature hyperparameter.

We use a VICReg-style variance-covariance regularizer to prevent predicted latent evidence repre-

sentations from collapsing (Bardes et al., 2022):

$$\mathcal{L}_{\text{sig}} = \frac{1}{d} \sum_{j=1}^d \max(0, 1 - \text{std}_j(\hat{\mathbf{Z}}^v))^2 + \frac{1}{d(d-1)} \sum_{i \neq j} \mathbf{C}_{ij}^2, \quad (8)$$

where $\hat{\mathbf{Z}}^v = \{\hat{\mathbf{z}}_{s,k}^v \mid s \in \mathcal{E}, k \in \mathcal{Y}\}$ collects all type hypotheses for gold entity spans, including the null hypothesis but excluding O spans; std_j is the per-dimension standard deviation, and $\mathbf{C}_{ij}$ is the (i, j) -th entry of the sample covariance matrix $\mathbf{C}$. The first term penalizes dimensional collapse through a squared hinge, and the second decorrelates dimensions.

The gate-weighted reverse alignment loss mirrors Eq. (7):

$$\mathcal{L}'_{\text{ira},g} = \frac{1}{|\mathcal{E}|} \sum_{s \in \mathcal{E}} \text{sg}(g_s) \|\hat{\mathbf{z}}_{s,y_s}^t - \text{sg}(\mathbf{z}_{\text{span},s})\|_2^2.$$

We stop gradients through g_s in auxiliary reverse losses so that the gate cannot reduce these losses by simply closing the reverse path. The gate-weighted reverse contrastive loss is

$$\mathcal{L}'_{\text{ico},g} = -\frac{1}{|\mathcal{E}|} \sum_{s \in \mathcal{E}} \text{sg}(g_s) \log \frac{\exp(s_{s,y_s}^{\text{rev}}/\tau)}{\sum_{k \in \mathcal{Y}} \exp(s_{s,k}^{\text{rev}}/\tau)}.$$

The reverse anti-collapse loss $\mathcal{L}'_{\text{sig}}$ follows Eq. (8) applied to predicted textual evidence representations $\hat{\mathbf{Z}}^t = \{\hat{\mathbf{z}}_{s,k}^t \mid s \in \mathcal{E}, k \in \mathcal{Y}\}$. Because anti-collapse regularization prevents predictor collapse rather than supervising evidence matching, we do not gate the reverse anti-collapse term. If a mini-batch contains no gold entity spans, we skip the auxiliary verification losses and optimize only $\mathcal{L}_{\text{task}}$.

C Additional Statistical Interpretation

Hypothesis-testing setup. Let $\mathcal{Y} = \{y_0\} \cup \mathcal{Y}^+$, where y_0 denotes the O/null class, $K = |\mathcal{Y}^+|$, and $C = |\mathcal{Y}| = K + 1$. For a candidate span u , let $Y_u \in \mathcal{Y}$ be its gold label. For each non-O type $k \in \mathcal{Y}^+$, MNER can be interpreted as a binary hypothesis test:

$$H_0^{u,k} : Y_u \neq k, \quad H_1^{u,k} : Y_u = k.$$

The null hypothesis is composite: the span may be a non-entity, or it may be an entity of another type. The O label therefore acts as the non-entity null class in final classification, while non-O labels correspond to type-specific alternatives.

Information-theoretic implication. Let Y^* denote the gold non-O entity type for spans in the gold entity set $\mathcal{E}$, and let $\mathbf{F} = \mathbf{s}^{\text{fwd}}$. The forward contrastive objective implies

$$I(Y^*; \mathbf{F}) \geq H(Y^*) - \mathcal{L}_{\text{ico}}.$$

Under a uniform prior over non-O entity types,

$$I(Y^*; \mathbf{F}) \geq \log |\mathcal{Y}^+| - \mathcal{L}_{\text{ico}}.$$

Thus, reducing $\mathcal{L}_{\text{ico}}$ both increases the soft evidence-ratio margin in Proposition 1 and makes the forward verification scores more informative about the entity type.

Score separation and testing error. For a threshold c_k , define the analytical test

$$\varphi_{u,k} = \mathbf{1}[T_{u,k} > c_k].$$

The type-I and type-II errors are

$$\begin{aligned} \alpha_k &= \Pr_{H_0^{u,k}}(T_{u,k} > c_k), \\ \beta_k &= \Pr_{H_1^{u,k}}(T_{u,k} \leq c_k). \end{aligned}$$

Let $\mu_{0,k} = \mathbb{E}[T_{u,k} \mid H_0^{u,k}]$, $\mu_{1,k} = \mathbb{E}[T_{u,k} \mid H_1^{u,k}]$, and $\Delta_k = \mu_{1,k} - \mu_{0,k}$. A larger Δ_k means that the verification statistic is better separated under the alternative and the null.

Proposition 3 (Score separation reduces testing error). *Suppose $T_{u,k}$ is sub-Gaussian with variance proxy σ^2 under both $H_0^{u,k}$ and $H_1^{u,k}$, and suppose $\Delta_k > 0$. If an oracle midpoint threshold is chosen as*

$$c_k = \frac{\mu_{0,k} + \mu_{1,k}}{2}, \quad (9)$$

then

$$\alpha_k \leq \exp\left(-\frac{\Delta_k^2}{8\sigma^2}\right), \quad \beta_k \leq \exp\left(-\frac{\Delta_k^2}{8\sigma^2}\right).$$

This analytical concentration result characterizes how an achieved increase in Δ_k improves the error exponent under the sub-Gaussian assumption. The contrastive loss $\mathcal{L}_{\text{ico}}$ promotes relative score separation, $\mathcal{L}_{\text{ira}}$ encourages predicted evidence to match observed evidence, and $\mathcal{L}_{\text{sig}}$ reduces degenerate cases in which type-conditioned evidence predictions become indistinguishable.

Multiple span-type tests and decoding. MNER requires many simultaneous span-type decisions:

$$\mathcal{H} = \{H_0^{u,k} : u \in \mathcal{S}(\mathbf{x}), k \in \mathcal{Y}^+\}.$$

At inference, the implementation selects the best non-O type $k_u^* = \arg \max_{k \in \mathcal{Y}^+} o_{u,k}$ and retains span u only if the entity-vs-null margin $m_u = o_{u,k_u^*} - o_{u,0}$ exceeds a development-tuned threshold δ . Overlapping predicted spans are then resolved greedily by confidence. This margin supplies the empirical rejection threshold used by the decoder. Future work could add calibration data to convert one-vs-rest verification statistics into empirical or conformal p -values under the composite null.

D Proofs

D.1 Proof of Proposition 1

For a gold entity span u with label $y \in \mathcal{Y}^+$, the forward contrastive loss is

$$\ell_u^{\text{fwd}} = -\log \frac{\exp(F_{u,y}/\tau)}{\sum_{k \in \mathcal{Y}} \exp(F_{u,k}/\tau)}.$$

Let

$$a_k = \frac{F_{u,k}}{\tau}.$$

Then

$$\ell_u^{\text{fwd}} = -\log \frac{\exp(a_y)}{\exp(a_y) + \sum_{\ell \in \mathcal{Y} \setminus \{y\}} \exp(a_\ell)}.$$

Rearranging,

$$\ell_u^{\text{fwd}} = \log \left(1 + \sum_{\ell \in \mathcal{Y} \setminus \{y\}} \exp(a_\ell - a_y) \right). \quad (10)$$

By definition,

$$\Lambda_{u,y}^{\text{fwd}} = a_y - \log \frac{1}{C-1} \sum_{\ell \in \mathcal{Y} \setminus \{y\}} \exp(a_\ell). \quad (11)$$

Exponentiating the negative of Eq. (11), we obtain

$$\begin{aligned} \exp(-\Lambda_{u,y}^{\text{fwd}}) &= \exp(-a_y) \cdot \\ &\quad \frac{1}{C-1} \sum_{\ell \in \mathcal{Y} \setminus \{y\}} \exp(a_\ell). \end{aligned}$$

Therefore,

$$\sum_{\ell \in \mathcal{Y} \setminus \{y\}} \exp(a_\ell - a_y) = (C-1) \exp(-\Lambda_{u,y}^{\text{fwd}}). \quad (12)$$

Substituting Eq. (12) into Eq. (10) gives Eq. (4). Finally, the derivative with respect to $\Lambda_{u,y}^{\text{fwd}}$ is

$$\frac{\partial \ell_u^{\text{fwd}}}{\partial \Lambda_{u,y}^{\text{fwd}}} = -\frac{(C-1) \exp(-\Lambda_{u,y}^{\text{fwd}})}{1 + (C-1) \exp(-\Lambda_{u,y}^{\text{fwd}})} < 0.$$

Thus, the instance-wise loss is strictly decreasing in $\Lambda_{u,y}^{\text{fwd}}$. This proves Proposition 1.

The evidence-ratio interpretation operates at the surrogate level. Treating $\exp(F_{u,k}/\tau)$ as an unnormalized evidence likelihood makes the second term in Eq. (11) a smooth mixture approximation of composite-null evidence over competing hypotheses; accordingly, $\Lambda_{u,y}^{\text{fwd}}$ is a logit-level evidence-ratio surrogate.

D.2 Proof of the Information-Theoretic Corollary

Let Y^* denote the gold non-O entity type for spans in $\mathcal{E}$, so $Y^* \in \mathcal{Y}^+$. Let $\mathbf{F} = \mathbf{s}^{\text{fwd}}$. Define the full softmax distribution over $\mathcal{Y}$ as

$$q_k(\mathbf{F}) = \frac{\exp(\mathbf{F}_k/\tau)}{\sum_{j \in \mathcal{Y}} \exp(\mathbf{F}_j/\tau)}.$$

The implemented evidence contrastive loss can be written as

$$\mathcal{L}_{\text{ico}} = \mathbb{E}_{Y^*, \mathbf{F}} [-\log q_{Y^*}(\mathbf{F})].$$

Although the denominator includes the O/null class, the positive label is always non-O. Normalize the predictive distribution over non-O labels:

$$Q_+(\mathbf{F}) = \sum_{k \in \mathcal{Y}^+} q_k(\mathbf{F}),$$

$$\bar{q}_k(\mathbf{F}) = \frac{q_k(\mathbf{F})}{Q_+(\mathbf{F})}, \quad k \in \mathcal{Y}^+.$$

Since $Q_+(\mathbf{F}) \leq 1$, for $Y^* \in \mathcal{Y}^+$ we have

$$\begin{aligned} -\log q_{Y^*}(\mathbf{F}) &= -\log \bar{q}_{Y^*}(\mathbf{F}) \\ &\quad -\log Q_+(\mathbf{F}) \\ &\geq -\log \bar{q}_{Y^*}(\mathbf{F}). \end{aligned}$$

Taking expectations gives

$$\mathcal{L}_{\text{ico}} \geq \mathbb{E}_{Y^*, \mathbf{F}} [-\log \bar{q}_{Y^*}(\mathbf{F})]. \quad (13)$$

The right-hand side is the cross-entropy of the normalized non-O classifier. By the standard cross-entropy decomposition,

$$\begin{aligned} &\mathbb{E}_{Y^*, \mathbf{F}} [-\log \bar{q}_{Y^*}(\mathbf{F})] \\ &= H(Y^* | \mathbf{F}) \\ &\quad + \mathbb{E}_{\mathbf{F}} \left[D_{\text{KL}}(p(Y^* | \mathbf{F}) \parallel \bar{q}(\cdot | \mathbf{F})) \right] \\ &\geq H(Y^* | \mathbf{F}). \end{aligned} \quad (14)$$

Combining Eq. (13) and Eq. (14),

$$H(Y^* | \mathbf{F}) \leq \mathcal{L}_{\text{ico}}.$$

Therefore,

$$\begin{aligned} I(Y^*; \mathbf{F}) &= H(Y^*) - H(Y^* | \mathbf{F}) \\ &\geq H(Y^*) - \mathcal{L}_{\text{ico}}. \end{aligned}$$

If the non-O entity types are uniformly distributed, then

$$H(Y^*) = \log |\mathcal{Y}^+| = \log K,$$

and hence

$$I(Y^*; \mathbf{F}) \geq \log |\mathcal{Y}^+| - \mathcal{L}_{\text{ico}}.$$

This proves the information-theoretic corollary. The same normalization argument applies to reverse scores when the corresponding reverse contrastive objective is considered without reliability weighting. The implemented $\mathcal{L}'_{\text{ico},g}$ uses stop-gradient gate weights as a reliability-aware training variant, and should not be read as changing the mutual-information identity itself.

D.3 Proof of Proposition 3

Fix a type k and omit k from the notation for clarity. Let

$$\begin{aligned} \mu_0 &= \mathbb{E}[T | H_0], \quad \mu_1 = \mathbb{E}[T | H_1], \\ \Delta &= \mu_1 - \mu_0 > 0. \end{aligned}$$

Assume T is sub-Gaussian with variance proxy σ^2 under both hypotheses. That is, for any $t > 0$,

$$\Pr_{H_0}(T - \mu_0 \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right), \quad (15)$$

and

$$\Pr_{H_1}(\mu_1 - T \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \quad (16)$$

Choose the oracle midpoint threshold

$$c = \frac{\mu_0 + \mu_1}{2} = \mu_0 + \frac{\Delta}{2} = \mu_1 - \frac{\Delta}{2}.$$

The type-I error is

$$\alpha = \Pr_{H_0}(T > c) = \Pr_{H_0}\left(T - \mu_0 > \frac{\Delta}{2}\right).$$

By Eq. (15),

$$\alpha \leq \exp\left(-\frac{(\Delta/2)^2}{2\sigma^2}\right) = \exp\left(-\frac{\Delta^2}{8\sigma^2}\right).$$

Similarly, the type-II error is

$$\beta = \Pr_{H_1}(T \leq c) = \Pr_{H_1}\left(\mu_1 - T \geq \frac{\Delta}{2}\right).$$

By Eq. (16),

$$\beta \leq \exp\left(-\frac{(\Delta/2)^2}{2\sigma^2}\right) = \exp\left(-\frac{\Delta^2}{8\sigma^2}\right).$$

This proves Proposition 3.

This proof uses an oracle midpoint threshold that depends on μ_0 and μ_1 . The actual implementation instead tunes an entity-vs-null margin on the development set. Therefore, Proposition 3 should be interpreted as a concentration-based analytical explanation of why larger separation is beneficial, not as a statement that the implemented decoder achieves the exact bound.

A related threshold choice gives an explicit significance-level interpretation. For a desired type-I upper bound α_0 , choose

$$c_{\alpha_0} = \mu_0 + \sigma\sqrt{2\log(1/\alpha_0)}.$$

Then

$$\Pr_{H_0}(T > c_{\alpha_0}) \leq \alpha_0.$$

Under H_1 ,

$$\begin{aligned} \beta &= \Pr_{H_1}(T \leq c_{\alpha_0}) \\ &= \Pr_{H_1}\left(\mu_1 - T \geq \Delta - \sigma\sqrt{2\log(1/\alpha_0)}\right). \end{aligned}$$

If

$$\Delta > \sigma\sqrt{2\log(1/\alpha_0)},$$

then

$$\beta \leq \exp\left(-\frac{\left(\Delta - \sigma\sqrt{2\log(1/\alpha_0)}\right)^2}{2\sigma^2}\right).$$

This shows that increasing the separation Δ improves power at a fixed type-I error level under the same concentration assumption.

D.4 Proof of Proposition 2

Let

$$F = F_{u,k}, \quad G = G_{u,k} = g_u R_{u,k}.$$

The bidirectional statistic is

$$T_\rho = F + \rho G.$$

Define the forward separation and gated reverse separation as

$$\begin{aligned} \Delta_f &= \mathbb{E}[F | H_1] - \mathbb{E}[F | H_0], \\ \Delta_g &= \mathbb{E}[G | H_1] - \mathbb{E}[G | H_0]. \end{aligned}$$

The separation of the combined statistic is

$$\Delta(\rho) = \mathbb{E}[T_\rho | H_1] - \mathbb{E}[T_\rho | H_0] = \Delta_f + \rho\Delta_g.$$

Assume F and G are independent sub-Gaussian statistics under both hypotheses, with variance proxies σ_f^2 and σ_g^2 . Then T_ρ is sub-Gaussian with variance proxy

$$\sigma^2(\rho) = \sigma_f^2 + \rho^2\sigma_g^2.$$

From Proposition 3, the testing error exponent is controlled by the signal-to-noise ratio

$$J(\rho) = \frac{\Delta(\rho)^2}{\sigma^2(\rho)} = \frac{(\Delta_f + \rho\Delta_g)^2}{\sigma_f^2 + \rho^2\sigma_g^2}.$$

The forward-only statistic corresponds to $\rho = 0$:

$$J(0) = \frac{\Delta_f^2}{\sigma_f^2}.$$

The derivative of $J(\rho)$ at $\rho = 0$ is

$$J'(0) = \frac{2\Delta_f\Delta_g}{\sigma_f^2}.$$

If $\Delta_f > 0$ and $\Delta_g > 0$, then $J'(0) > 0$. Therefore, there exists a sufficiently small $\rho > 0$ such that

$$J(\rho) > J(0).$$

Since the concentration bound in Proposition 3 improves as $J(\rho)$ increases, the bidirectional statistic can yield a more powerful test than the forward-only statistic when gated reverse verification provides positive additional separation. This proves the first part of Proposition 2.

Next suppose reverse evidence contributes no useful separation under unreliable visual evidence, so $\Delta_g \approx 0$. Under the same independence assumption,

$$J(\rho) \approx \frac{\Delta_f^2}{\sigma_f^2 + \rho^2\sigma_g^2} \leq \frac{\Delta_f^2}{\sigma_f^2} = J(0).$$

Thus, adding uninformative reverse evidence can increase variance without improving separation. If the gate learns $g_u \approx 0$ for unreliable images, then

$G = g_u R_{u,k} \approx 0$, and the statistic effectively reverts to the forward-only case. This proves the reliability-gating claim under the stated assumptions.

In practice, F and G need not be exactly independent. A more general local condition can be stated by writing the variance proxy as

$$\sigma^2(\rho) = \sigma_f^2 + 2\rho\kappa + \rho^2\sigma_g^2,$$

where κ represents a covariance proxy between F and G . Then

$$J(\rho) = \frac{(\Delta_f + \rho\Delta_g)^2}{\sigma_f^2 + 2\rho\kappa + \rho^2\sigma_g^2}.$$

A direct derivative calculation gives

$$J'(0) = \frac{2\Delta_f(\Delta_g\sigma_f^2 - \Delta_f\kappa)}{\sigma_f^4}.$$

Therefore, when $\Delta_f > 0$, a sufficient condition for small positive ρ to improve the signal-to-noise ratio is

$$\Delta_g\sigma_f^2 > \Delta_f\kappa.$$

This correlated-case condition makes explicit that reverse verification is beneficial when its additional separation is large enough to offset any covariance-induced increase in effective variance.

E Grounded and Fine-Grained Extension Results

For GMNER and FMNERG, we follow the official evaluation protocols and report the main grounded metric together with the corresponding text-only entity recognition metric and the EEG metric. These comparisons evaluate the extensibility of TCHV on grounded benchmarks.

F Experimental Details

F.1 Hyperparameters

Table 11 lists all hyperparameters.

For TCHV-BERT, the text encoder is replaced with frozen BERT-base, while all other hyperparameters in Table 11 remain unchanged.

F.2 Computing Environment

All experiments were conducted on a Linux server equipped with an Intel Xeon Platinum 8580 CPU and approximately 503 GiB system memory. Model training was performed on a single NVIDIA H200 NVL GPU with about 140 GB memory. The

software environment was managed with Conda. We used Python 3.11.9 and PyTorch 2.6.0+cu124. The NVIDIA driver version was 570.195.03; the CUDA Runtime version was 12.8, while PyTorch was built against CUDA 12.4. The main dependencies included Transformers 5.4.0, NumPy 1.26.4, Pillow 12.1.0, and scikit-learn 1.8.0. FP16 mixed-precision training was enabled for the standard MNER experiments, and VinVL region features were used for the GMNER/FMNERG experiments.

F.3 Verification Diagnostic Protocol and Matched Controls

For dataset d , let $\mathcal{G}_d = \{u : y_u \in \mathcal{Y}^+\}$ denote all gold non-O test spans. Given their gold boundaries, the forward scores predict among the four entity types:

$$\hat{y}_u^{\text{fwd}} = \arg \max_{k \in \mathcal{Y}^+} s_{u,k}^{\text{fwd}},$$

$$\text{Acc}_{\text{fwd}}(d) = \frac{1}{|\mathcal{G}_d|} \sum_{u \in \mathcal{G}_d} \mathbb{I}[\hat{y}_u^{\text{fwd}} = y_u].$$

With learned type embeddings, the forward scores select the annotated type for 4547/5181 spans on Twitter-2015 and 1269/1338 spans on Twitter-2017, corresponding to 87.76% and 94.84% accuracy. For the counterfactual, a fixed cyclic permutation π over $\mathcal{Y}^+$ satisfies $\pi(k) \neq k$ and substitutes $\mathbf{e}_{\pi(k)}$ when evaluating hypothesis k . Both conditions use the same checkpoint, text, images, comparators, and scoring modules; only the candidate-type-embedding association changes. The cyclic condition yields 235/5181 (4.54%) and 28/1338 (2.09%), supporting the type-specific correspondence between candidate hypotheses and expected evidence.

For Table 4(b), all four systems use the same frozen RoBERTa-base and ViT-Base/16 backbones and the same training and decoding pipeline. Direct fusion removes candidate-type hypotheses, type-conditioned evidence construction, forward/reverse verification, and verification scores, and directly fuses the text and image representations. The three TCHV variants retain both verification paths and the expected-versus-observed comparison, changing the conditioning operator to concatenation, cross-attention, or AdaLN.

F.4 Gate Initialization Sensitivity

We evaluate four initial gate priors on Twitter-2017. The gate remains trainable in every setting, and each sensitivity run is trained for 10 epochs. The

Table 9: Results on GMNER. We report P/R/F1 (%) for GMNER, MNER, and EEG, following H-Index (Yu et al., 2023). Within this table, best and second-best scores are highlighted in bold and underline, respectively.

Method	GMNER			MNER			EEG		
	P	R	F1	P	R	F1	P	R	F1
<i>Text-only methods</i>									
HBiLSTM-CRF-None	43.56	40.69	42.07	78.80	72.61	75.58	49.17	45.92	47.49
BERT-None	42.18	43.76	42.96	77.26	77.41	77.30	46.76	48.52	47.63
BERT-CRF-None	42.73	44.88	43.78	77.23	78.64	77.93	46.92	49.28	48.07
BARTNER-None	44.61	45.04	44.82	79.67	79.98	79.83	48.77	49.23	48.99
<i>Text+Image methods</i>									
GVATT-RCNN-EVG	49.36	47.80	48.57	78.21	74.39	76.26	54.19	52.48	53.32
UMT-RCNN-EVG	49.16	51.48	50.29	77.89	79.28	78.58	53.55	56.08	54.78
UMT-VinVL-EVG	50.15	52.52	51.31	77.89	79.28	78.58	54.35	56.91	55.60
UMGF-VinVL-EVG	51.62	51.72	51.67	79.02	78.64	78.83	55.68	55.80	55.74
ITA-VinVL-EVG	52.37	50.77	51.56	80.40	78.37	79.37	56.57	54.84	55.69
BARTMNER-VinVL-EVG	52.47	52.43	52.45	<u>80.65</u>	80.14	80.39	55.68	55.63	55.66
H-Index	<u>56.16</u>	<u>56.67</u>	<u>56.41</u>	<u>79.37</u>	<u>80.10</u>	<u>79.73</u>	<u>60.90</u>	61.46	<u>61.18</u>
TCHV + Extension	62.23	57.41	59.73	83.59	77.11	<u>80.22</u>	65.64	<u>60.56</u>	63.00

Table 10: Results on FMNERG. We report P/R/F1 (%) for FMNERG, FMNER, and EEG, following TIGER (Wang et al., 2023). Within this table, best and second-best scores are highlighted in bold and underline, respectively.

Method	FMNERG			FMNER			EEG		
	P	R	F1	P	R	F1	P	R	F1
<i>Text-only methods</i>									
HBiLSTM-CRF-None	34.86	32.38	33.57	62.31	56.55	59.29	49.25	43.27	46.07
BERT-None	33.28	34.28	33.77	58.91	60.05	59.47	46.27	47.65	46.94
BERT-CRF-None	34.41	35.51	34.95	60.06	61.38	60.72	46.93	48.44	47.67
T5-Paraphrase-None	37.38	37.29	37.33	64.83	65.32	65.07	49.03	48.91	48.97
<i>Text+Image methods</i>									
GVATT-VinVL-EVG	42.02	38.75	40.32	63.08	57.85	60.35	55.27	53.46	54.35
UMT-VinVL-EVG	40.67	41.99	41.32	61.24	62.01	61.63	53.58	55.32	54.43
UMGF-VinVL-EVG	41.73	42.11	41.92	61.68	61.90	61.79	54.51	55.00	54.75
ITA-VinVL-EVG	43.05	42.51	42.78	63.80	62.64	63.21	57.63	56.90	57.26
H-Index	46.83	46.28	46.55	65.25	64.45	64.84	60.82	60.10	60.46
MMT5-VinVL-EVG	45.35	45.08	45.21	<u>66.46</u>	66.77	<u>66.61</u>	58.35	58.01	58.18
TIGER	<u>47.57</u>	<u>46.85</u>	<u>47.20</u>	64.43	<u>65.40</u>	64.91	<u>62.44</u>	61.49	<u>61.96</u>
TCHV + Extension	53.05	48.44	50.64	70.39	64.22	67.16	66.74	<u>60.93</u>	63.70

default prior $\sigma(\theta_0) = 0.8808$ gives the highest F1 among the tested initializations. The variation across priors supports the gate’s role as a learned reliability controller.

F.5 Offline VLM Knowledge Sources

For standard MNER, TCHV[†] reuses the artifacts released by PGIM (Li et al., 2023a), including ChatGPT-generated entity candidates and the contextual reasoning provided for Twitter-2017. For GMNER and FMNERG, it uses offline BLIP-2 captions (Li et al., 2023b), injected through a lightweight gated cross-attention module. These resources are prepared before downstream training, which updates the TCHV task modules. The enhanced results show that the verification framework can incorporate existing offline knowledge.

F.6 Data Preprocessing

Sentences are tokenized with the RoBERTa tokenizer (maximum 128 sub-word tokens). We maintain a word-to-token alignment map to recover span boundaries in token space. Candidate spans are enumerated for all (i, j) pairs with $j - i + 1 \leq L$; spans extending beyond the sentence are discarded. The maximum number of candidate spans per sentence is capped at 192.

Images are resized to 224×224 pixels using the standard ViT preprocessing pipeline, yielding $N = 14 \times 14 = 196$ patch tokens plus one [CLS] token. Together, these steps align word-level spans with encoder inputs and provide a fixed visual-token grid for type-conditioned verification.

Table 11: Full hyperparameter configuration.

Hyperparameter	Value
<i>Model</i>	
Text encoder	RoBERTa-base (125M)
Image encoder	ViT-Base/16 (86M)
Shared dimension d	384
Type embedding dimension d_e	128
Predictor hidden dimension d_p	512
Predictor layers M	4
FFN expansion factor	2
Comparator hidden dimension	256
Classifier hidden dimension	256
Width embedding dimension d_w	32
Maximum span length L	4
Maximum sequence length	128
<i>Training</i>	
Epochs	50
Batch size	16
Encoder learning rate	0 (frozen)
New module learning rate	1×10^{-4}
Weight decay	0.01
Warmup ratio	0.06
LR scheduler	Cosine decay
Max gradient norm	1.0
Mixed precision	FP16
EMA decay	0.995
<i>Loss</i>	
α (forward alignment)	1.0
α' (reverse alignment)	1.0
β (forward contrastive)	0.5
β' (reverse contrastive)	0.5
γ (forward anti-collapse)	0.05
γ' (reverse anti-collapse)	0.05
Temperature τ	0.07
Label smoothing	0.02
Non-entity weight w_0	0.25

Table 12: Twitter-2017 gate-initialization sensitivity after 10 training epochs. The gate is trainable in every run.

Initial prior $\sigma(\theta_0)$	Test F1
0.5000	88.80
0.7311	88.91
0.8808 (default)	89.85
0.9526	88.95

F.7 Margin Threshold Tuning

At inference, we apply a margin-based decoding strategy. For each candidate span u , we first select $k_u^* = \arg \max_{k \in \mathcal{Y}^+} o_{u,k}$ and compute $m_u = o_{u,k_u^*} - o_{u,0}$ using the pre-softmax logits $\mathbf{o}_u$, i.e., the margin between the highest entity-type logit and the non-entity logit. A span is predicted as an entity only if $m_u > \delta$, where δ is selected on the dev set by grid search over $[0.0, 2.0]$ with step 0.01. Conflicts are resolved by greedy non-overlapping selection, preferring higher-margin spans.

F.8 Extension to GMNER and FMNERG

For GMNER and FMNERG, we keep the span-level TCHV backbone unchanged and replace image-patch visual evidence with proposal-level visual evidence because grounded prediction requires region-level outputs (Yu et al., 2023). The proposal encoder provides region features, region boxes (Zhang et al., 2021), and region masks, which serve as the observed visual evidence for grounding. The type-conditioned latent prediction and verification modules are reused to produce entity-type verification scores for each candidate span.

To support grounded prediction, we attach a lightweight region pointer on top of TCHV. The pointer scores candidate proposal regions using the span representation and type-aware verification features. Therefore, the grounded extension can be viewed as adding a region selection branch to the same span-level hypothesis-verification architecture.

For GMNER (Yu et al., 2023), the model predicts entity-type-region triples using the standard coarse entity type space. For FMNERG (Wang et al., 2023), the type hypothesis bank and classifier are expanded to the fine-grained label space, while the same grounding head is retained. Coarse-to-fine calibration and auxiliary knowledge injection, when used, act as lightweight enhancements to fine-grained classification rather than changes to the TCHV backbone; this is consistent with recent fine-grained grounded MNERG studies (Wang et al., 2024). All grounded results are evaluated using the official protocols of the corresponding benchmarks.