MVCCLF: Gradient Surgery-Driven Multi-View Contrastive Learning for Multimodal NER

1st Kai Zhuo
Zhejiang Sci-Tech University
Hangzhou, China
2024220603118@mails.zstu.edu.cn

2nd Yifan Huo
Zhejiang Sci-Tech University
Hangzhou, China
2024010602002@mails.zstu.edu.cn

3rd Junhong Zheng*
Zhejiang Sci-Tech University
Hangzhou, China
zjhist@zstu.edu.cn

4th Lili He
Zhejiang Sci-Tech University
Hangzhou, China
llhe@zju.edu.cn

***Abstract*—Multimodal Named Entity Recognition (MNER) significantly improves entity recognition performance by integrating textual and visual information from social media posts. However, existing methods still face challenges such as imprecise text-image semantic alignment, visual noise interference, and multi-task optimization conflicts, resulting in insufficient fine-grained cross-modal interaction and unstable training. To address these issues, this paper proposes a Gradient Surgery-Driven Multi-View Cross-Modal Contrastive Learning Framework (MVCCLF). The framework enables fine-grained cross-modal alignment via a coarse-to-fine multi-view contrastive learning mechanism, while introducing dynamic weight scheduling and a dual-layer gradient surgery mechanism to effectively mitigate gradient conflicts between the primary NER task and auxiliary contrastive learning. Experimental results on the Twitter15 and Twitter17 datasets demonstrate that MVCCLF achieves F1 scores of 75.55% and 87.31%, respectively, demonstrating competitive performance on both datasets. Ablation studies and sensitivity analysis further validate the effectiveness of each component and the robustness of the framework, providing an efficient fine-grained fusion solution for MNER in social media scenarios.**

***Index Terms*—Contrastive learning, Multimodal Named Entity Recognition, Gradient Surgery, Transformer**

I. Introduction

Named Entity Recognition (NER) aims to identify and classify specific entities such as persons (PER), locations (LOC), and organizations (ORG) from unstructured text [1]. Traditional NER methods face severe challenges when processing social media content, where informal expressions, grammatical irregularities, and limited contextual information often lead to significant declines in recognition accuracy. To address this, MNER introduces image information to assist textual understanding, providing a novel pathway to improve entity recognition performance in social media scenarios [2].

As illustrated in Figure 1, MNER leverages cross-modal context to resolve entity ambiguity. The text "Handsome Rob after a fish dinner" alone leaves "Rob" type-ambiguous (PER vs. animal), but integrating the accompanying image of a cat enables precise classification as MISC.

* Corresponding author: Junhong Zheng.

Fig. 1. Text : Handsome Rob after a fish dinner

Currently, fusion methods based on cross-modal attention mechanisms have become the dominant paradigm in MNER research [3]. However, the problem of inconsistent semantic alignment quality between text and images in social media remains largely unresolved, as users frequently share images that are weakly related or even completely unrelated to the accompanying text, thereby introducing substantial visual noise that interferes with entity recognition [4]. Although existing studies have proposed various noise-reduction strategies—such as auxiliary entity span detection [5], dynamic gating mechanisms for acquiring multi-scale visual features [6], graph convolutional networks as substitutes for Transformer-based fusion modules [7], and sentence-level contrastive learning [8]—two critical limitations persist: first, in terms of feature interaction granularity, existing approaches either focus on global image-sentence alignment or merely incorporate simple visual object features, failing to establish explicit, fine-grained semantic associations from text tokens to local image patches; second, in terms of optimization strategy, naively combining contrastive learning objectives with sequence labeling tasks through simple weighted joint training readily triggers gradient conflicts, causing mutual interference in multi-task optimization and hindering convergence to an optimal solution [9]. Meanwhile, recent advances in contrastive learning have demonstrated remarkable progress in cross-modal represen-

tation learning, with methods such as CLIP [10] showing that aligning visual and textual representations via contrastive objectives enables strong zero-shot transfer capabilities. Nevertheless, directly applying such global-level contrastive learning to MNER is insufficient, because entity recognition inherently requires fine-grained token-level understanding rather than mere sentence-level alignment.

To tackle the above challenges, we introduce three complementary contrastive learning objectives operating at different granularities. Global-View Contrastive Learning (GVCL) aligns sentence-level textual representations with image-level visual features to establish coarse-grained semantic correspondence; Region-Token Ranking Contrastive Learning (RTRCL) learns fine-grained alignment between text tokens and image patches via a Top-K ranking mechanism, enabling entity-aware cross-modal interaction; and Self-Augmented Visual Contrastive Learning (SAVCL) enhances the robustness of visual representations through self-supervised augmentation. Drawing on the idea of PCGrad [9](Projected Conflicting Gradients), we incorporate a gradient surgery mechanism to resolve inherent optimization conflicts in multi-task learning. Additionally, we adopt a three-stage progressive training strategy to ensure stability and effective optimization throughout the entire training process. Ultimately, we propose a Gradient Surgery-Driven Multi-View Cross-Modal Contrastive Learning Framework (MVCCLF) for multimodal named entity recognition, which effectively addresses the aforementioned issues.

The main contributions of this paper are summarized as follows:

- We propose MVCCLF, which achieves multi-view cross-modal alignment in MNER—from sentence-level semantics to token-patch fine-grained correspondence and visual robustness—through the complementary integration of GVCL, RTRCL, and SAVCL.
- We introduce a dual-layer gradient surgery mechanism and design a three-stage progressive training strategy with dynamic loss weight scheduling, which effectively mitigates multi-task optimization conflicts and ensures stable convergence.
- Comprehensive experiments and sensitivity analyses conducted on two benchmark datasets demonstrate the effectiveness and superiority of our proposed method.

II. RELATED WORK

A. Multimodal Named Entity Recognition (MNER)

MNER extends traditional NER by integrating visual information from accompanying images in text. Moon et al. [2] pioneered the fusion of global visual features via attention mechanisms, demonstrating that images can provide valuable disambiguation signals for entity recognition in social media. Subsequent research has developed more complex cross-modal fusion strategies: object detection-based methods, represented by Zhang et al. [11] and Sun et al. [12], leverage pre-trained detectors to extract region-level features, offering targeted visual guidance for entities. However, these methods suffer from significant limitations: semantic misalignment between visual features and NER objectives; single-granularity attention struggles to effectively distinguish entity-relevant areas from background noise; and simple weighted multi-task optimization overlooks gradient conflicts, resulting in unstable training.

Recent works address challenges like text-image semantic misalignment and visual noise. RiVEG [13] leverages LLMs to reformulate Grounded MNER as a joint MNER-Visual Entailment-Visual Grounding pipeline, yet fine-grained token-patch alignment and multi-task gradient conflicts remain unresolved. Multi-granularity approaches such as M3S [14] and MGICL [15] attempt to capture cross-modal interactions from global to fine-grained levels; BFCL [16] combines bottleneck fusion with contrastive learning to align representations while reducing redundancy. Nevertheless, existing methods predominantly rely on sentence-level positive-negative sampling, failing to capture token-level visual evidence—e.g., "Washington" as a person name corresponding to a portrait or as a location corresponding to the White House—thus limiting fine-grained entity disambiguation. Moreover, current works generally overlook self-augmented views of the same image to learn visual invariance representations, which is particularly crucial for handling social media images with varying illumination, angles, and quality.

B. Vision-Language Pre-training and Contrastive Learning

Vision-language pre-trained models, trained on large-scale image-text pairs via contrastive learning, provide powerful cross-modal alignment capabilities for downstream multimodal tasks. CLIP [10], trained on 400 million image-text pairs, aligns textual and visual information into a unified semantic space using cross-modal contrastive learning. Its Vision Transformer encoder produces 196 patch tokens, offering finer-grained representations more suitable for token-to-patch alignment compared to ResNet's convolutional features [17]. Recent MNER works have begun exploring CLIP applications: MAF [8] leverages CLIP's global [CLS] token to enhance textual representations but underutilizes patch-level features for fine-grained alignment; ICKA [18] aligns CLIP's visual encoder with the NER task through instruction tuning, yet remains confined to image-level interactions. Multi-granularity contrastive learning further extends the semantic hierarchy of cross-modal alignment. ALBEF [19] conducts contrastive learning at three granularities—image-text, region-word, and feature-feature—to capture coarse-to-fine cross-modal correspondences. MGIPS-CL [20] proposes implicit positive sample mining by computing relevance between text and all images in a batch, dynamically selecting Top-K most relevant images as positives to mitigate mismatches in original pairings. However, its contrastive learning is still performed at the image level, failing to establish fine-grained correspondences between tokens and patches.

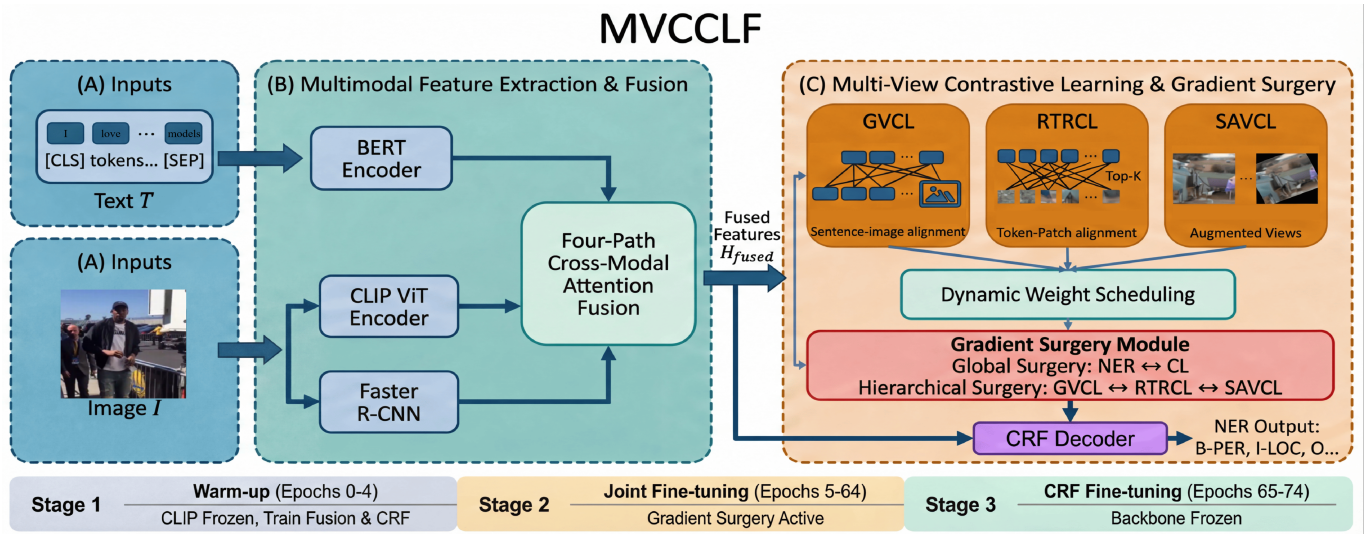

Fig. 2. Overall architecture of MVCCLF. The framework integrates (A) text-image inputs, (B) multimodal feature extraction and four-path cross-modal fusion using BERT, CLIP ViT, and Faster R-CNN encoders, and (C) multi-view contrastive learning with gradient surgery.

C. Multi-Task Learning and Gradient Conflict Resolution

Multi-task learning aims to improve generalization by jointly optimizing multiple related tasks through shared representations. When tasks have conflicting objectives, naive joint optimization via gradient averaging can lead to suboptimal solutions. Yu et al. [9] observed on multi-task benchmarks that gradients from different tasks often exhibit negative cosine similarity, causing parameter updates to cancel each other out and resulting in slow convergence or performance degradation. Consequently, PCGrad introduces gradient surgery as a principled approach to resolve conflicts in multi-task learning: when conflict is detected, it projects one task's gradient onto the orthogonal plane of another, removing destructive components while preserving constructive updates.

In the context of MNER, multi-task frameworks combining sequence labeling with auxiliary objectives face similar gradient conflict challenges. For instance, existing methods such as UMT [5] and BFCL [16] adopt simple weighted loss summation, overlooking conflicts among task gradients. Moreover, when models encounter multi-view contrastive learning, different contrastive losses may also conflict: for example, global alignment tends to pull sentence and whole-image representations closer, while local alignment pulls token and patch representations closer, producing inconsistent gradient directions on shared encoders.

III. METHOD

This section presents the proposed MVCCLF, which addresses the imprecise alignment between textual and visual information in MNER tasks. By constructing three complementary contrastive learning perspectives—global, local, and self-augmented—the framework achieves fine-grained cross-modal alignment. It further integrates a gradient surgery-driven three-stage progressive training strategy to enable effective modeling of multimodal information. The overall architecture is illustrated in Figure 2.

A. Multimodal Feature Representations

1) Feature Extraction: The text encoder is based on the pre-trained BERT-base-cased model [21]. The visual encoder employs a dual-path design to extract global semantic features and local object features, respectively. Specifically, the CLIP visual encoder utilizes the pre-trained ViT-B/16 model [22] to extract global semantic features of the image. The object detector adopts the pre-trained Faster R-CNN [23] to capture object information in the image and extract the Top-3 object region features with the highest confidence scores.

2) Cross-Modal Feature Fusion: We adopt a four-path fusion strategy to achieve deep interaction between textual and visual features through cross-modal attention mechanisms. The cross-modal attention function CMA(Q, K, V) performs standard multi-head attention operations. The outputs of the four paths can be uniformly expressed as:

$$H^{(k)} = \mathrm{CMA}(H^{\mathrm{text}}, V^{(k)}, V^{(k)}), \quad k \in K = \{\mathrm{self}, \mathrm{global}, \mathrm{patch}, \mathrm{obj}\} \tag{1}$$

$$\text{where } V^{(k)} = \begin{cases} H^{\mathrm{text}}, & k = \mathrm{self} \\ H^{\mathrm{global}}, & k = \mathrm{global} \\ H^{\mathrm{patch}}, & k = \mathrm{patch} \\ H^{\mathrm{obj}}, & k = \mathrm{obj} \end{cases} \tag{2}$$

After concatenation of the four-path features, a linear projection layer is applied for fusion:

$$H^{\mathrm{fused}} = \mathrm{ReLU}\big(W_f \cdot [H^{(\mathrm{self})}; H^{(\mathrm{global})}; H^{(\mathrm{patch})}; H^{(\mathrm{obj})}] + b_f\big) \tag{3}$$

where $W_f \in \mathbb{R}^{768\times 3072}$ and $[\cdot;\cdot]$ denotes concatenation.

B. Multi-View Cross-Modal Contrastive Learning Mechanism

Traditional MNER methods struggle to capture fine-grained cross-modal semantics via sentence-level alignment. We propose a tri-perspective (global-local-self-augmented) cross-modal contrastive learning mechanism with dynamic weighting for synergistic optimization, ensuring global alignment, local detail matching, and enhanced visual robustness.

1) Global-View Contrastive Learning(GVCL): GVCL aims to establish a global semantic bridge between text and images, providing macro-level semantic constraints for fine-grained alignment. Given a batch of size B, the [CLS] representations of text $\{t_i\}_{i=1}^{B}$ and images $\{v_i\}_{i=1}^{B}$ serve as positive pairs ($i = j$), while other samples in the batch act as negatives ($i \neq j$). The normalized similarity matrix is first computed as:

$$S_{ij} = \frac{t_i \cdot v_j}{\|t_i\|\|v_j\|\tau_g} \tag{4}$$

where τ_g is the temperature hyperparameter. The bidirectional cross-entropy loss is then calculated as:

$$\mathcal{L}_{t2i} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ij})} \tag{5}$$

The image-to-text direction is computed symmetrically, yielding the final GVCL loss:

$$\mathcal{L}_{GVCL} = \frac{1}{2}(\mathcal{L}_{t2i} + \mathcal{L}_{i2t}). \tag{6}$$

When constructing training batches, we ensure diversity in entity types (PER, LOC, ORG, MISC) and prioritize hard negatives with similar entity types but different specific entities, thereby adapting GVCL to the NER task.

2) Region-Token Ranking Contrastive Learning(RTRCL): RTRCL achieves token-level fine-grained cross-modal alignment by dynamically selecting the Top-K most relevant image patches for each text token, capturing precise visual semantic information through ranking-based selection and feature pooling. Given a text token sequence $T = \{t_i\}_{i=1}^{|T|}$ and an image patch sequence $\{p_j\}_{j=1}^{|P|}$, for each token t_i, the Top-K most similar patches (based on cosine similarity) form the positive set $\mathcal{P}_i^+$, while the Bottom-K form the negative set $\mathcal{P}_i^-$ (with $K = 3$). The selected patch features are aggregated via average pooling:

$$\bar{p_i^+} = \frac{1}{K}\sum_{p_j\in\mathcal{P}_i^+} p_j, \quad \bar{p_i^-} = \frac{1}{K}\sum_{p_j\in\mathcal{P}_i^-} p_j \tag{7}$$

The contrastive loss is computed using the logsigmoid form of InfoNCE:

$$\mathcal{L}_{t2i} = -\frac{1}{|T|}\sum_{i=1}^{|T|}\log\sigma\left(\frac{\text{sim}(t_i,\bar{p_i^+}) - \text{sim}(t_i,\bar{p_i^-})}{\tau_l}\right) \tag{8}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ_l is the temperature hyperparameter. The image-to-text direction is computed symmetrically, yielding:

$$\mathcal{L}_{\text{RTRCL}} = \frac{1}{2}(\mathcal{L}_{t2i} + \mathcal{L}_{i2t}). \tag{9}$$

3) Self-Augmented Visual Contrastive Learning(SAVCL): SAVCL constructs multiple views of each image through data augmentation to strengthen the model's learning of visual invariance, enabling stable feature extraction even in scenarios with varying image quality (illumination, angles, resolution). For each image V, an augmented view V^{aug} is generated. The original view and its augmentation serve as a positive pair, while views from different images act as negatives. The SAVCL loss is defined as:

$$\mathcal{L}_{SAVCL} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(z_i^{\text{orig}}, z_i^{\text{aug}})/\tau_a)}{\sum_{j=1,j\neq i}^{2B}\exp(\text{sim}(z_i^{\text{orig}}, z_j)/\tau_a)} \tag{10}$$

where $z_i^{\text{orig}} and z_i^{\text{aug}}$ denote the features of the original view and its corresponding augmented view for the i-th sample in the batch , τ_a is the temperature hyperparameter, and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. The loss is computed symmetrically in both directions (orig→aug and aug→orig). The denominator excludes self-pairs and corresponding views ($j \neq i, i+B$), ensuring only same-source views are treated as positives among the 2B views.

4) Dynamic Weight Scheduling Mechanism: Different training batches exhibit varying degrees of reliance on the three contrastive learning objectives: simple sentences rely more on global alignment, sentences with complex entities require finer-grained alignment, and entities of the MISC category place greater emphasis on visual consistency. To adaptively balance these three contrastive learning tasks with the main NER task, we propose a dynamic weight scheduling mechanism. The gating weight learner takes the average representation of text [CLS] features within the batch, denoted as $\bar{h}_{cls}^{text}$, as contextual input, and learns normalized weights through a two-layer fully connected network:

$$(\alpha,\beta,\gamma) = \text{Softmax}\left(W_2\cdot\text{ReLU}\left(W_1\cdot\bar{h}_{cls}^{text} + b_1\right) + b_2\right) \tag{11}$$

where $W_1 \in \mathbb{R}^{384\times768}$, $W_2 \in \mathbb{R}^{3\times384}$, and the output weights satisfy $\alpha+\beta+\gamma = 1$. The three contrastive losses are first combined via these dynamic weights, then modulated by an outer temporal weight to obtain the multi-view contrastive loss:

$$\mathcal{L}_{CL} = \lambda(t)\Big[\alpha\mathcal{L}_{GVCL} + \beta\mathcal{L}_{RTRCL} + \gamma\mathcal{L}_{SAVCL}\Big] \tag{12}$$

The outer weight $\lambda(t)$ controls the overall strength of the contrastive losses and adopts a warmup strategy that dynamically adjusts with training steps:

$$\lambda(t) = \lambda_{\max}\cdot\left[\frac{2}{1+e^{-6t/T}} - 1\right], \tag{13}$$

where $\lambda_{\max}$ is the maximum weight and T is the warmup step count. This design ensures that $\lambda(t) \approx 0$ at the early training stage ($t \approx 0$), allowing the model to focus on foundational learning of the NER task, with $\lambda(t)$ gradually increasing to introduce contrastive constraints as training progresses. Additionally, the sequence labeling task employs a CRF [24]

as the decoding layer. The NER loss is defined as the negative log-likelihood:

$$\mathcal{L}_{NER} = -\log P(Y \mid X, I) \tag{14}$$

where $P(Y \mid X, I)$ is computed by the CRF layer, incorporating emission scores and a learnable transition matrix. Based on the above, the final overall loss function is:

$$\mathcal{L} = \mathcal{L}_{NER} + \lambda(t) \cdot [\alpha \mathcal{L}_{GVCL} + \beta \mathcal{L}_{RTRCL} + \gamma \mathcal{L}_{SAVCL}] \tag{15}$$

C. Gradient Surgery-Driven Training Strategy

In multi-task learning, conflicting gradient directions across tasks can lead to mutual interference during optimization. Since the MVCCLF framework simultaneously optimizes four loss functions—NER loss, GVCL loss, RTRCL loss, and SAVCL loss—this paper employs Gradient Surgery techniques to resolve gradient conflicts between the NER task and the three contrastive learning tasks (integrated with a three-stage progressive strategy to mitigate catastrophic forgetting). The core principle is: when the cosine similarity between gradient vectors of two tasks is detected to be negative <0, the conflicting gradient is projected onto the orthogonal space of the other gradient, retaining only the non-conflicting components for parameter updates.

1) Gradient Conflict Detection and Projection: Let $\nabla \mathcal{L}_{NER}$, $\nabla \mathcal{L}_{GVCL}$, $\nabla \mathcal{L}_{RTRCL}$, and $\nabla \mathcal{L}_{SAVCL}$ denote the gradients of the four loss functions with respect to the shared parameters. A conflict exists when the cosine similarity between two gradients is negative:

$$\text{conflict}(\nabla \mathcal{L}_i, \nabla \mathcal{L}_j) = \mathbb{1}[\cos(\nabla \mathcal{L}_i, \nabla \mathcal{L}_j) < 0] \tag{16}$$

Upon conflict detection, the conflicting gradient is projected onto the orthogonal space of the other gradient, retaining only the perpendicular component:

$$\nabla \mathcal{L}_i' = \nabla \mathcal{L}_i - \frac{\nabla \mathcal{L}_i \cdot \nabla \mathcal{L}_j}{\|\nabla \mathcal{L}_j\|^2} \nabla \mathcal{L}_j \tag{17}$$

This operation eliminates destructive gradient components, enabling synergistic multi-task optimization.

2) Global Gradient Surgery: Global gradient surgery treats the primary NER task and all contrastive learning tasks as two independent task groups. We compute the gradients of the NER loss and the total contrastive loss $\mathcal{L}_{CL}$ with respect to shared parameters. Upon conflict detection, we apply a unidirectional projection: the NER gradient is projected onto the orthogonal space of the contrastive gradient (removing conflicting components), while the contrastive gradient remains unchanged:

$$\nabla \mathcal{L}_{NER}' = \nabla \mathcal{L}_{NER} - \frac{\nabla \mathcal{L}_{NER} \cdot \nabla \mathcal{L}_{CL}}{\|\nabla \mathcal{L}_{CL}\|^2} \nabla \mathcal{L}_{CL} \tag{18}$$

The final aggregated gradient is:

$$g_{\text{final}} = \nabla \mathcal{L}_{NER}' + \nabla \mathcal{L}_{CL} \tag{19}$$

3) Hierarchical Gradient Surgery: To address internal conflicts among the three contrastive learning tasks within the MVCCLF framework, we further propose a hierarchical gradient surgery strategy, applying targeted gradient projections to different parameter groups. The model parameters are divided into three functional groups: BERT parameter group (affected by NER, GVCL, and RTRCL); CLIP parameter group (affected by GVCL, RTRCL, and SAVCL); and Fusion layer parameter group (affected by all tasks).

For each parameter group, we apply the PCGrad projection strategy to handle gradient conflicts. Let $G = \{g_1, g_2, \ldots, g_k\}$ be the set of gradients for a given parameter group. For each gradient g_i, we iterate over all other gradients g_j ($j \neq i$); upon conflict detection, g_i is projected onto the orthogonal space of g_j:

$$g_i' = g_i - \frac{g_i \cdot g_j}{\|g_j\|^2} g_j \tag{20}$$

The final aggregated gradient for the group is the average of all modified gradients:

$$g_{\text{final}} = \frac{1}{k} \sum_{i=1}^{k} g_i' \tag{21}$$

The specific scheduling of global and hierarchical surgery across training stages, along with layer freezing and learning rate configurations, is provided in Section 4.1.

IV. EXPERIMENT

A. Experimental Setup

1) Datasets: To ensure fair comparison with prior works, we conduct experiments on two widely-used public benchmarks for MNER: Twitter15 [3] and Twitter17 [25]. Twitter15 contains 8,257 tweets with 12,800 annotated entities, while Twitter17 includes 4,819 tweets with 8,724 entities. Following the data splits adopted by Wu et al. [26] and Yu et al. [5], we divide Twitter15 into 4,000/1,000/3,257 for training/validation/test sets, and Twitter17 into 3,373/723/723.

2) Hyperparameter Settings: Our method is implemented in PyTorch and trained on an NVIDIA RTX A6000 GPU. We use identical hyperparameter configurations across both datasets. The batch size is set to 32 for training and 16 for validation and testing, with a maximum text length of 128 tokens. All encoder outputs are unified to 768 dimensions (consistent with the configuration described in Section 3.1). The model comprises a single Transformer layer with 12 attention heads and a dropout rate of 0.1. Temperature parameters are configured as follows: GVCL $\tau = 0.07$, RTRCL $\tau = 0.05$, and SAVCL $\tau = 0.06$.

3) Three-Stage Training Strategy: We adopt a progressive three-stage training pipeline to fully leverage the advantages of multimodal fusion:

- Stage 1(Epochs 0–4): Warm-up stage. The CLIP backbone is frozen; only the cross-modal attention and CRF layers are trained. Learning rates: BERT 3×10^{-5}, cross-modal attention 1×10^{-4}, CRF 3×10^{-5}.

TABLE I

RESULTS COMPARISON OF BASELINES ON TWITTER15 AND TWITTER17. PERFORMANCE COMPARISON OF TEXT-ONLY AND MULTIMODAL METHODS ON TWO BENCHMARK DATASETS. BOLD VALUES INDICATE THE BEST PERFORMANCE, AND UNDERLINED VALUES INDICATE OUR METHOD'S RESULTS.

Methods	**Twitter15**							**Twitter17**						
	Single Type (F1)				Overall			Single Type (F1)				Overall		
	PER	LOC	ORG	MISC	P	R	F1	PER	LOC	ORG	MISC	P	R	F1
Text														
CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37
HBiLSTM-CRF	82.34	76.83	51.59	32.52	70.32	68.05	69.17	87.91	78.57	76.67	59.32	82.69	78.16	80.37
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
Text+Image														
AdapCoAtt	85.28	80.64	59.39	38.88	69.87	74.59	72.15	90.20	82.97	82.67	64.83	85.13	83.20	84.10
UMT	85.24	81.58	63.03	39.45	71.84	74.61	73.20	91.56	84.73	82.24	70.10	85.08	85.27	85.18
MEGA	-	-	-	-	70.35	74.58	72.35	-	-	-	-	84.03	84.75	84.39
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	-	-	-	-	73.87	**76.82**	75.32	-	-	-	-	85.84	87.93	86.87
BFCL	85.60	81.77	63.81	40.30	74.02	75.07	74.54	91.17	86.43	83.97	66.67	85.99	85.42	85.70
M3S	86.05	81.32	62.97	41.36	**74.9**	75.14	75.03	92.73	84.81	82.49	69.53	**86.93**	85.21	86.06
CRISP	85.59	80.57	61.16	38.24	71.92	74.57	73.22	93.10	84.66	85.54	68.24	86.76	87.34	87.05
ICAK	87.01	83.85	65.87	48.28	72.36	78.75	75.42	93.99	87.24	86.24	75.76	85.13	**89.19**	87.12
MVCCLF	86.89	83.23	64.62	41.19	74.70	76.42	**75.55**	93.15	85.53	86.69	68.90	86.75	87.87	**87.31**

- Stage 2(Epochs 5–64): Joint fine-tuning stage. The top layers of CLIP (Blocks 10–11) are unfrozen, with an initial learning rate of 6×10^{-6} linearly warmed up to 1.5×10^{-5}. Gradient surgery is performed in two phases: the first 30 epochs address global conflicts between NER and MVCCLF, while the next 30 epochs implement hierarchical balancing among GVCL, RTRCL, and SAVCL. Learning rates: BERT 1×10^{-5}, cross-modal attention 5×10^{-5}, CRF 5×10^{-4}, GVCL 5×10^{-5}, RTRCL 4×10^{-5}, SAVCL 6×10^{-5}.
- Stage 3(Epochs 65–74): Backbone fully frozen; only the CRF layer is fine-tuned with a learning rate of 2×10^{-4}. Gradient surgery is disabled.

4) Data Augmentation: We apply random image cropping (scale 0.85–1.0), rotation (±10°), and grayscale conversion (probability 0.1).

5) Evaluation Metrics: Following standard practice in prior MNER methods, we report Precision (Pre), Recall (Rec), and macro-averaged F1 (F1) scores on the test set. Additionally, we provide per-entity-type F1 scores for PER, LOC, ORG, and MISC on both Twitter datasets.

B. Overall Comparison Results

1) Baseline Methods Introduction: To validate the effectiveness of our proposed MVCCLF method, we compare it against a total of 11 baseline methods, including both text-only and multimodal approaches. The baselines are as follows:

- CNN-BiLSTM-CRF [27]: A classic CNN-BiLSTM encoder with CRF decoder for sequence labeling.
- BERT-CRF [21]: BERT-pretrained encoder for text with CRF for sequence labeling.
- HBiLSTM-CRF [24]: Hierarchical BiLSTM for text encoding combined with CRF decoding.
- UMT [5]: Unified multimodal Transformer framework with auxiliary entity span detection to mitigate irrelevant visual noise.
- MAF [8]: A contrastive learning framework that debiases visual features to suppress irrelevant visual interference.
- MEGA [7]: GCN-based dual-graph alignment method capturing correlations between entities and visual objects for cross-modal alignment.
- HVPNet [6]: ResNet-Transformer hybrid using visual representations as prefixes with dynamic gating for multi-scale visual features.
- BFCL [16]: Transformer-based model incorporating bottleneck fusion and contrastive learning to optimize cross-modal representations and reduce redundancy.
- M3S [14]: Scene graph-driven multi-granularity multi-task framework aligning object and entity views for enhanced visual-text utilization.
- CRISP [28]: Cross-modal integration framework employing the Surprisingly Popular Algorithm for text-image fusion in MNER.
- ICKA [18]:Instruction-based knowledge alignment framework that captures cross-modal entity semantics via explicit instructions.

2) Experimental Results Analysis: We refer to the experimental results reported in the original papers of the baseline methods and conduct a comprehensive comparison with 11 representative approaches on the Twitter15 and Twitter17 datasets (including text-only unimodal methods such as BERT-CRF, and multimodal methods such as UMT, MAF, MEGA, HVPNet, etc.). The results are shown in Table I.

Among text-only unimodal methods, BERT-based models achieve F1 scores of 71.81% and 83.44% on the two

TABLE II
ABLATION STUDY RESULTS

Model Variant	Twitter15							Twitter17						
	Single Type (F1)				Overall			Single Type (F1)				Overall		
	PER	LOC	ORG	MISC	P	R	F1	PER	LOC	ORG	MISC	P	R	F1
w/o GVCL	86.31	83.28	62.62	40.69	74.45	75.67	75.05	92.28	85.83	84.73	68.71	86.15	86.65	86.40
w/o RTRCL	85.76	83.04	63.01	40.04	73.90	75.24	74.57	92.82	81.64	85.44	67.73	85.03	87.32	86.16
w/o SAVCL	86.52	83.17	63.53	40.12	74.12	76.16	75.13	92.97	83.35	85.71	67.91	85.46	87.73	86.58
w/o Gradient Surgery	85.72	82.91	63.23	40.76	73.32	76.11	74.69	92.18	84.69	85.68	68.34	84.98	87.67	86.30
MVCCLF	86.89	83.23	64.62	41.19	74.70	76.42	**75.55**	93.15	85.53	86.69	68.90	86.75	87.87	**87.31**

datasets, respectively, significantly outperforming traditional CNN/LSTM baselines. This highlights the superiority of pre-trained encoders for NER tasks on social media text and explains why recent multimodal approaches commonly adopt BERT-style structures as the textual backbone. In the multimodal setting, methods that employ Faster R-CNN for pre-detection of objects (e.g., HVPNet at 75.32%/86.87% and M3S at 75.03%/86.06%) substantially outperform simple attention-based fusion through complex integration and multi-scale visual features. Regarding patch-level fine-grained interaction, while MEGA and MAF introduce patch information, they fail to fully exploit dynamic token-patch alignment. In contrast, our MVCCLF's RTRCL module dynamically selects the Top-K=3 most relevant patches for each text token, enabling more precise fine-grained alignment. Under the contrastive learning paradigm, MAF and BFCL are limited to sentence-image-level or single-view contrast, neglecting multi-granularity semantic mining and thus constraining performance.

In summary, through its multi-view contrastive learning framework (GVCL, RTRCL, SAVCL) and gradient surgery-driven three-stage training strategy, MVCCLF effectively addresses conflicts and imbalances in cross-modal fusion. It achieves the best F1 scores of 75.55% and 87.31% on Twitter15 and Twitter17, respectively, demonstrating clear advantages in fine-grained alignment, multi-view consistency, and training stability.

C. The Ablation Study

In this section, we conduct ablation experiments on the Twitter15 and Twitter17 datasets to verify the effectiveness of each component in the MVCCLF framework. We design the following ablation variants to evaluate the contributions of key modules:

- w/o GVCL: denotes the removal of the Global View Contrastive Learning module.
- w/o RTRCL: denotes the removal of the Region-Token Ranking Contrastive Learning module.
- w/o SAVCL: denotes the removal of the Self-Augmented View Contrastive Learning module.
- w/o Gradient Surgery: denotes the removal of the gradient surgery mechanism.

As shown in Table II, all components are beneficial for MNER, validating the effectiveness of multi-view contrastive learning. Removing GVCL, RTRCL, or SAVCL all leads to a decrease in F1 score, highlighting the complementary contributions of the three modules. Among them, the removal of RTRCL has the most significant impact, indicating that Top-K cross-modal alignment is crucial; removing GVCL impairs global semantic understanding; and removing SAVCL reduces visual robustness. After removing the gradient surgery mechanism, performance drops significantly, proving that there exists notable gradient conflict between the NER task and contrastive learning losses.

D. Hyperparameter Experiments

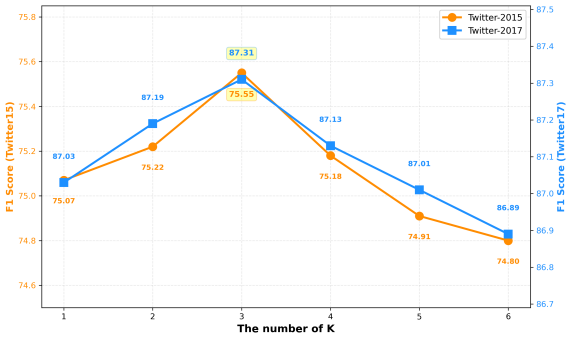

Fig. 3. Sensitivity analysis of Top-K values in RTRCL.

1) Sensitivity Analysis of Top-K Value in RTRCL: The hyperparameter K of the RTRCL module controls the number of patches selected per token, influencing the fine-grained alignment performance. As shown in Figure 3, sensitivity experiments on the Twitter15 and Twitter17 datasets demonstrate that the F1 score peaks at K=3 (75.55% and 87.31%, respectively). An excessively small K value results in insufficient visual semantic capture, while an overly large K value introduces noisy patches, diminishing the effectiveness of contrastive learning.

2) Sensitivity Analysis of Dynamic Weight $\lambda_{\max}$: The dynamic weight $\lambda_{\max}$ controls the contrastive learning loss intensity, balancing the main NER task and auxiliary tasks. As shown in Figure 4, performance peaks at $\lambda_{\max} = 0.15$. Smaller values provide insufficient regularization, causing the model to degrade into relying solely on textual features, while larger values interfere with the sequence labeling task. This finding

is consistently validated on both Twitter15 and Twitter17, confirming that $\lambda_{\max}$=0.15 achieves the optimal balance.

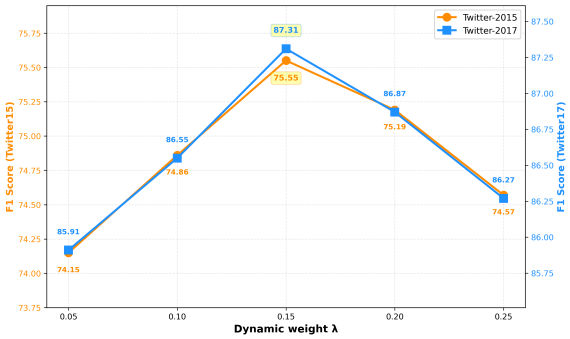

Fig. 4. Sensitivity analysis of dynamic weight $\lambda_{\max}$

V. CONCLUSION

This paper proposes MVCCLF to address imprecise cross-modal alignment, gradient conflicts, and visual noise in MNER. Through GVCL, RTRCL, and SAVCL, we achieve fine-grained alignment from global to token-patch level while enhancing visual invariance. Dynamic weight scheduling with dual-level gradient surgery mitigates multi-task optimization conflicts, ensuring synergy between primary and auxiliary objectives. Experiments on Twitter15 and Twitter17 validate the model's effectiveness and module designs. Future work will explore leveraging multimodal LLMs for deep semantic extraction and zero-shot MNER.

ACKNOWLEDGMENT

This work was supported by the "Pioneer" R&D Program of Zhejiang (Grant No. 2026LDC01024(XC)) and the "Pioneer" R&D Program of Zhejiang (Grant No. 2025C01088).

REFERENCES

[1] Li, J., Sun, A., Han, J., & Li, C. (2022). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50-70.

[2] Moon, S., Neves, L., & Carvalho, V. (2018). Multimodal named entity recognition for short social media posts. In Proceedings of NAACL-HLT (pp. 852-860).

[3] Zhang, Q., Fu, J., Liu, X., & Huang, X. (2018). Adaptive co-attention network for named entity recognition in tweets. In Proceedings of the AAAI Conference on Artificial Intelligence, 32.

[4] Vempala, A., & Preoţiuc-Pietro, D. (2019). Categorizing and inferring the relationship between the text and image of Twitter posts. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2830-2840).

[5] Yu, J., Jiang, J., Yang, L., & Xia, R. (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 (pp. 3342-3352).

[6] X. Chen et al., "Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction," arXiv preprint arXiv:2205.03521, 2022.

[7] C. Zheng, J. Feng, Z. Fu, Y. Cai, Q. Li, and T. Wang, "Multimodal relation extraction with efficient graph alignment," in *Proc. 29th ACM* *Int. Conf. Multimedia*, Oct. 2021, pp. 5298–5306.

[8] B. Xu, S. Huang, C. Sha, and H. Wang, "MAF: A general matching and alignment framework for multimodal named entity recognition," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, Feb. 2022, pp. 1215–1223.

[9] Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. Advances in Neural Information Processing Systems, 33, 5824-5836.

[10] Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.

[11] Zhang, D., Wei, S., Li, S., Wu, H., Zhu, Q., & Zhou, G. (2021). Multi-modal graph fusion for named entity recognition with targeted visual guidance. In Proceedings of the AAAI Conference on Artificial Intelligence, 35, 14347-14355.

[12] Sun, L., Wang, J., Zhang, K., Su, Y., & Weng, F. (2021). RpBERT: A Text-image Relation Propagation-based BERT Model for Multimodal NER. Proceedings of the AAAI Conference on Artificial Intelligence, 35(15), 13860-13868.

[13] J. Li et al., "LLMs as Bridges: Reformulating Grounded Multimodal Named Entity Recognition," in Findings of the Association for Computational Linguistics: ACL 2024, pp. 1302–1318, 2024.

[14] Wang, J., Yang, Y., Liu, K., Zhu, Z., & Liu, X. (2022). M3S: Scene graph driven multi-granularity multi-task learning for multi-modal NER. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 111-120.

[15] Liu, P., Wang, G., Li, H., Liu, J., Ren, Y., Zhu, H., & Sun, L. (2024). Multi-granularity cross-modal representation learning for named entity recognition on social media. Information Processing and Management, 61, 103546.

[16] Wang, P., Chen, X., Shang, Z., & Ke, W. (2023). Multimodal named entity recognition with bottleneck fusion and contrastive learning. IEICE Transactions on Information and Systems, E106.D(4), 545-555.

[17] A. K. Monsefi, H. Sedghi, A. Nagrani, A. Zisserman, and H. Jégou, "DetailCLIP: Detail-Oriented CLIP for Fine-Grained Tasks," *arXiv preprint* arXiv:2409.06809, 2024.

[18] Zeng, Q., Yuan, M., Wan, J., Wang, K., Shi, N., Che, Q., & Liu, B. (2024). ICKA: An instruction construction and knowledge alignment framework for multimodal named entity recognition. Expert Systems with Applications, 255, 124867.

[19] Li, J., Selvaraju, R. R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems, 34, 9694-9705.

[20] Wei, P., Chen, S., Wen, Q., Hu, Q., Zeng, B., & Feng, G. (2025). MGIPS-CL: Multimodal NER with multi-granularity interweave and implicit positive sample contrastive learning. Applied Soft Computing, 187, 114298.

[21] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

[22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[23] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. NeurIPS.

[24] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., San Diego, CA, USA, K. Knight, A. Ne-nkova, and O. Rambow, Eds., 2016, pp. 260–270.

[25] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "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)* , 2018,pp. 1990–1999.

[26] Wu, Z., Zheng, C., Cai, Y., Chen, J., Leung, H., & Li, Q. (2020). Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts. *Proceedings of the 28th ACM International Conference on Multimedia*.ings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp.1990–1999.

[27] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, in:Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1064–1074.

[28] H. Liu, X. Xin, J. Song, W. Peng, CRISP: A Cross-modal Integration Framework Based on the Surprisingly Popular Algorithm for Multimodal Named Entity Recognition, Neurocomputing 614 (2025) 128792.