A Hybrid Multimodal Architecture for Named Entity Recognition on Social Media Using Domain-Specific Language Modelling and Token-Level Gated Visual Fusion

Ayesha Shehzadi
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
ayesha.msds22seecs@seecs.edu.pk

Muhammad Moazam Fraz
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
moazam.fraz@seecs.edu.pk

Syed Imran Ali
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
imran.ali@seecs.edu.pk

Hafiz Syed Muhammad Bilal
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
bilal.ali@seecs.edu.pk

Fahad Ahmad Satti
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
fahad.satti@seecs.edu.pk

***Abstract*—Named entity recognition on social media is a markedly harder problem than its counterpart on formal text, owing to the informal vocabulary, creative abbreviations, and context-dependent entity references that characterise platforms such as Twitter. Although attaching an image to a tweet often carries the disambiguating information needed to resolve ambiguous entity mentions, most existing methods either discard the visual channel entirely or incorporate it in ways that introduce noise rather than signal. We present MSCMT_Hybrid, a hybrid architecture that encodes tweet text with BERTweet - a domain-specific language model pre-trained on 850 million tweets — and extracts global visual semantics with a frozen CLIP ViT-L/14 encoder. A learnable cross-modal projection maps the image embedding into the text representation space, after which a token-level sigmoid gate selectively modulates visual influence at each token position. A Conditional Random Field decoder enforces globally valid BIO label sequences. We evaluate the model on a merged corpus combining TWITTER-2015 and TWITTER-2017, totalling approximately 12,090 tweet-image pairs. A systematic grid search over learning rate and batch size yields a final weighted F1 of 75.97% - a competitive result relative to published state-of-the-art methods on this task, achieved with a substantially simpler cross-modal fusion mechanism than prior Transformer-based approaches. We additionally conduct a cross-dataset generalisation analysis and ablation study that identify the individual contribution of each architectural component, and document a systematic failure mode of unified vision-language models when adapted to token-level sequence labelling, providing an empirical motivation for the hybrid design.**

Index Terms-Multimodal Named Entity Recognition, Social Media NLP, BERTweet, CLIP ViT-L/14, Token-Level Gated Fusion, Conditional Random Field, Vision-Language Models, Twitter NER.

I. Introduction

Social media platforms have become primary channels through which people report breaking events, share opinions, and discuss public figures. Twitter alone generates hundreds of millions of posts daily, each potentially containing references to persons, locations, organisations, and other named entities. Extracting these entities automatically - a task known as Named Entity Recognition (NER) - underpins a wide range of downstream applications, from event detection and knowledge graph construction to misinformation tracking and public health surveillance.

What makes NER on social media particularly challenging is the nature of the content itself. Tweets are short, grammatically informal, and often rely on shared cultural context that a purely text-based model cannot access. Consider a tweet reading "Jordan drops 40 at the United Center"- whether "Jordan" refers to a basketball player, a brand, or a country cannot be resolved from the text alone. An attached photograph of a basketball game, however, makes the answer immediate. This observation motivates the broader area of Multimodal Named Entity Recognition (MNER), which jointly processes text and images to improve entity identification in social media posts [1].

Early MNER approaches established the task's viability by showing that visual attention over image regions improves NER compared to text-only baselines [2], [3]. Subsequent

work introduced cross-modal Transformer layers and graph-based entity linking to more deeply integrate the two modalities [4]. The most recent systems - including MGCMT [5] - employ multi-granularity cross-modal fusion with object-level visual representations extracted by Mask RCNN. Despite this architectural sophistication, a core question remains open: how much of the performance gain comes from the quality of the visual encoder, and how much from the complexity of the fusion mechanism?

This paper addresses that question directly. We propose MSCMT_HYBRID, a model built on two deliberate design choices. First, we replace prior text encoders - which typically used BERT-base trained on Wikipedia and BooksCorpus - with BERTweet [6], a RoBERTa-based model pre-trained on 850 million tweets. This shift aligns the language model's training distribution with the target domain. Second, we use CLIP ViT-L/14 [7] as our visual encoder, kept frozen to preserve the multimodal representations learned from 400 million image-text pairs, and introduce a lightweight token-level sigmoid gate to regulate how visual information flows into each token representation. The gate is conditioned on both the textual token embedding and the projected image representation, learning to amplify visual cues when relevant and suppress noise otherwise.

We additionally document an important negative finding. Before arriving at the hybrid design, we investigated whether end-to-end vision-language models - specifically CLIP [7] and BLIP-2 [8] - could be directly adapted to the MNER task. Both configurations stagnated at test accuracies of approximately 49-52% despite convergence of training loss. Our analysis indicates a structural mismatch between the global, sentence-level pre-training objectives of these models and the local, per-token prediction granularity required for sequence labelling. This finding provides a principled justification for the hybrid design and offers a replicable diagnostic observation for the community.

The contributions of this paper are as follows. We present a hybrid MNER architecture that combines domain-specific text encoding with frozen CLIP visual features through a token-level gating mechanism, achieving a weighted F1 score of 75.97% on the merged TWITTER-2015 and TWITTER-2017 benchmark - competitive with published state-of-the-art systems. We provide both empirical and conceptual analysis explaining why unified vision-language models underperform on token-level NER. Finally, we conduct a systematic grid search over training hyperparameters, demonstrating that the proposed architecture remains stable across a range of configurations.The complete training and inference procedure, including the three-phase optimisation strategy and Viterbi decoding, is formalised in Algorithm 1.We further provide a cross-dataset generalisation analysis and a two-condition ablation study that together identify the individual contributions of the text encoder and fusion mechanism, and empirically motivate the corpus merging strategy.

The remainder of the paper is structured as follows. Section II reviews related work. Section III describes the proposed architecture. Section IV presents experimental results and analysis. Section V concludes the paper.

II. RELATED WORK

A. NER as Structured Sequence Labelling

The foundations most relevant to our decoder were laid by architectures that treated NER as a structured prediction problem rather than a per-token classification problem [9]–[11]. The key insight - scoring entire label sequences jointly through a CRF rather than making independent token decisions - proved remarkably durable, and it carries forward directly into our own design. Contextual pre-training later shifted the bottleneck from architecture to representation quality: ELMo [12] and then BERT [13] delivered token embeddings that capture long-range syntactic and semantic dependencies, and the resulting BERT-CRF pipeline became the standard text-only baseline against which every subsequent MNER system is compared. On formal text this baseline is strong. On Twitter it is not, for reasons that go beyond domain shift in the ordinary sense: the tokeniser was simply not built for hashtags, abbreviations, or the compressed register of user-generated content, and no amount of fine-tuning fully compensates for that mismatch.

B. Multimodal NER and the Case for Selective Fusion

The multimodal direction grew from a single practical observation: a surprising fraction of the entity ambiguities that defeat text-only models dissolve the moment the accompanying image is examined [2], [3]. Both studies also surfaced a complication that has shaped every architecture since - a large proportion of tweet images are irrelevant to the entity mentions in the text, so unconditional fusion hurts as often as it helps. The field's answer has been gating: learn to open the visual channel when the image is informative and close it otherwise. Cross-modal Transformers pushed this further by replacing handcrafted gates with learned attention across modalities [4], [14]. The subsequent generation moved to object-level representations, arguing that disambiguation turns on specific detected objects rather than global scene content [5], [15]. Each step improved results and added complexity: the strongest published system requires three unimodal Transformer encoders, a Mask RCNN object detector, and two cross-modal interaction modules [5]. What none of these systems asks is whether the gains attributed to fusion design might partly reflect the ceiling imposed by a weak text encoder. Every one of them uses BERT-base - a model whose pre-training corpus shares almost nothing with tweet text. That question is precisely what motivates the design choices here.

C. Twitter-Specific Language Modelling

BERTweet [6] remains the most direct answer to the encoder problem: a RoBERTa-scale model pre-trained entirely on Twitter data, with a vocabulary built from the tweet corpus after platform-specific normalisation. The performance gap over BERT-base on Twitter NER is consistent and widest on the entity mentions that matter most - hashtag-format

Algorithm 1 MSCMT_Hybrid: Training and Inference

Input: Token sequence $\mathbf{X}$, Image $\mathbf{I}$, Dataset $\mathcal{D}$, Epochs T
Output: Label sequence $\mathbf{Y}^* = \{y_0, \ldots, y_M\}$

PHASE 1 - INITIALISATION

Load θ_{bert} from pre-trained BERTweet
Load θ_{clip} from CLIP ViT-L/14 **(freeze all weights)**
Randomly initialise $\mathbf{W}_p$, $\mathbf{W}_g$, $\mathbf{W}_c$, $\mathbf{A}$
Set $\text{lr}_{pre} = 3 \times 10^{-5}$ for $\{\theta_{bert}, \mathbf{W}_p\}$; $\text{lr}_{new} = 3 \times 10^{-4}$ for $\{\mathbf{W}_g, \mathbf{W}_c, \theta_{CRF}\}$
Initialise AdamW optimiser with weight decay
Build *WeightedRandomSampler* from inverse class frequencies

PHASE 2 - TRAINING LOOP

for epoch $= 1$ **to** T **do**
  for each mini-batch $(\mathbf{X}, \mathbf{I}, \mathbf{Y})$ from $\mathcal{D}$ **do**
    $\mathbf{H} \leftarrow \text{BERTweet}(\mathbf{X})$ ▷ $\mathbf{H} \in \mathbb{R}^{(M+1)\times 768}$
    $\mathbf{v} \leftarrow \text{CLIP}_{img}(\mathbf{I})$ ▷ $\mathbf{v} \in \mathbb{R}^{768}$
    $\mathbf{v}_{proj} \leftarrow \mathbf{W}_p \cdot \mathbf{v} + \mathbf{b}_p$ ▷ $\mathbf{v}_{proj} \in \mathbb{R}^{768}$
    for $i = 0$ **to** M **do**
      $\mathbf{g}_i \leftarrow \sigma(\mathbf{W}_g \cdot [\mathbf{h}_i; \mathbf{v}_{proj}] + \mathbf{b}_g)$ ▷ $\mathbf{g}_i \in (0,1)^{768}$
      $\text{fused}_i \leftarrow \text{Concat}(\mathbf{h}_i, \mathbf{g}_i \odot \mathbf{v}_{proj})$ ▷ $\mathbb{R}^{1536}$
      $\mathbf{e}_i \leftarrow \mathbf{W}_c \cdot \text{fused}_i + \mathbf{b}_c$ ▷ $\mathbf{W}_c \in \mathbb{R}^{9\times 1536}$
    $\mathcal{L} \leftarrow -\sum_t \log p(\mathbf{Y}_t \mid \mathbf{E}_t; \theta_{CRF})$
    $\Theta \leftarrow \Theta - \nabla_\Theta \mathcal{L}$ ▷ AdamW update
    Evaluate weighted F1 on validation set
    if no improvement for 3 consecutive epochs **then**
      break ▷ Early stopping

PHASE 3 - INFERENCE

Forward pass with frozen model, no gradient update
$\mathbf{Y}^* \leftarrow \text{Viterbi}(\mathbf{E}, \mathbf{A})$ ▷ *Globally optimal label sequence*
return $\mathbf{Y}^*$

names and abbreviated expressions that a standard tokeniser reduces to uninformative fragments. The broader finding that temporal and domain alignment of pre-training data matters more than model size [16] only sharpens the case. Despite this evidence, BERTweet has not appeared as the text encoder in any published MNER system. That absence is one of two gaps the proposed architecture fills.

D. Vision-Language Models and Structured Prediction

The failure of unified vision-language models on dense prediction tasks is not unique to NER. Studies in computer vision have shown that CLIP representations degrade on tasks requiring spatial localisation, because contrastive pre-training provides no gradient signal for per-pixel or per-token outputs [17], [18]. The same failure mode appeared in our preliminary experiments with CLIP and BLIP-2 adapted to MNER: training loss converged, but test accuracy stagnated near 50% with the model effectively predicting the majority O class throughout. The lesson is not that these models are unsuitable for multimodal tasks — it is that they should be used for what they were trained to do. CLIP's image encoder produces rich global visual embeddings that transfer well; its text encoder should not be asked to do token-level sequence labelling. MSCMT_Hybrid keeps exactly this division of labour, freezing CLIP for visual encoding and replacing everything else with components built for the target task.

E. Summary and Positioning

Table I summarises the key design choices of representative MNER systems across text encoder, visual encoder, fusion strategy, decoder, and evaluation datasets. Four observations emerge. First, no published MNER system has

adopted BERTweet as the text encoder despite its documented advantages on Twitter text. Second, CLIP ViT-L/14 has not been explored as a visual backbone for MNER, though it offers substantially richer semantic representations than ResNet alternatives. Third, token-level gated fusion has only appeared within complex multi-granularity pipelines, leaving open whether a simpler single-level gate is sufficient when encoder quality is high. Fourth, the failure of unified vision-language models on MNER has not been systematically documented in the published literature. The proposed MSCMT Hybrid architecture addresses all four gaps simultaneously, pairing BERTweet with a frozen CLIP encoder and bridging their embedding spaces through a lightweight token-level sigmoid gate, without requiring an object detector or multi-granularity fusion pipeline.

III. Proposed Methodology

A. Task Definition

Given a tweet represented as a token sequence

$$X = \{x_0, x_1, \ldots, x_M\}$$

and an accompanying image I, the goal is to assign a label $y_i \in \mathcal{Z}$ to each token, where $\mathcal{Z}$ denotes the BIO2 label set: {`B-PER`, `I-PER`, `B-LOC`, `I-LOC`, `B-ORG`, `I-ORG`, `B-OTHER`, `I-OTHER`, `O`}. The joint prediction over the full sequence is denoted as

$$Y = \{y_0, y_1, \ldots, y_M\}$$

B. Architecture Overview

MSCMT_Hybrid processes input through four sequential stages: (1) unimodal encoding, where text and image are independently encoded by pre-trained models; (2) cross-modal projection, mapping the image embedding into the textual representation space; (3) token-level gated fusion, selectively enriching token representations with visual information; and (4) structured decoding using a CRF layer to enforce globally valid label sequences. The overall architecture is illustrated in Fig. 1.

C. Text Encoder: BERTweet

The text encoder is `vinai/bertweet-base` [6], a 12-layer Transformer with hidden size 768, pre-trained on 850 million English tweets using the RoBERTa masked language modelling objective.

Given token sequence X, the encoder produces contextual representations:

$$H = \text{BERTweet}(X; \theta_{\text{bert}}) \in \mathbb{R}^{(M+1)\times 768}$$

where $M+1$ accounts for the prepended `[CLS]` token.

For tokens split into sub-words, the representation of the first sub-word is used as the token representation. BERTweet parameters are fine-tuned with a learning rate of 3×10^{-5}.

D. Visual Encoder: CLIP ViT-L/14

The visual encoder is `openai/clip-vit-L/14` [7], a Vision Transformer with patch size 14, processing images resized to 224×224 pixels. It was pre-trained using contrastive alignment over 400 million image-text pairs and is kept frozen during training:

$$v = \text{CLIP}_{\text{img}}(I; \theta_{\text{clip}}) \in \mathbb{R}^{768}$$

Freezing CLIP prevents catastrophic forgetting and reduces the number of trainable parameters.

E. Cross-Modal Projection and Broadcast

Although both BERTweet and CLIP produce 768-dimensional embeddings, their representation spaces differ. A learnable linear projection aligns the visual embedding with the textual space:

$$v_{\text{proj}} = W_p v + b_p, \quad W_p \in \mathbb{R}^{768\times 768}, \quad v_{\text{proj}} \in \mathbb{R}^{768}$$

The projected image vector is broadcast across all token positions:

$$V_{\text{proj}} \in \mathbb{R}^{(M+1)\times 768}$$

where each row is an identical copy of v_{proj}.

F. Token-Level Gated Fusion

For each token position i, a dimension-wise gate vector is computed:

$$g_i = \sigma(W_g[h_i; v_{\text{proj}}] + b_g), \quad W_g \in \mathbb{R}^{768\times 1536}$$

where $g_i \in (0,1)^{768}$ and $[\cdot;\cdot]$ denotes concatenation.

The gate is applied through element-wise multiplication:

$$\tilde{v}_i = g_i \odot v_{\text{proj}}$$

The fused representation is then formed via concatenation:

$$f_i^{\text{fused}} = \text{Concat}(h_i, \tilde{v}_i) \in \mathbb{R}^{1536}$$

where $\odot$ denotes element-wise multiplication.

This formulation preserves the original textual representation h_i without modification. Visual information enters only through the gated image component $\tilde{v}_i$, resulting in a 1536-dimensional fused vector. A dropout layer with rate of 0.1 is applied to the BERTweet output prior to fusion.

TABLE I. COMPARISON OF REPRESENTATIVE MNER SYSTEMS. T-2015 = TWITTER-2015; T-2017 = TWITTER-2017. PROPOSED MODEL HIGHLIGHTED.

Method	Year	Text Encoder	Visual Encoder	Fusion Strategy	Decoder	Dataset(s)
Moon et al. [1]	2018	BiLSTM	VGG-16	Modality-attention gate	CRF	T-2015, T-2017
Zhang et al. [3]	2018	BiLSTM	ResNet-50	Adaptive co-attention	CRF	T-2015, T-2017
Lu et al. [2]	2018	HBiLSTM	ResNet (regions)	Visual attention + gate	CRF	T-2015
UMT [4]	2020	BERT-base	ResNet (regions)	Cross-modal Transformer	CRF	T-2015, T-2017
OCSGA [14]	2020	BERT-base	Faster RCNN	Object-aware cross-attn	CRF	T-2015
ATTR-MMKG [15]	2021	BERT-base	ResNet + KG	Multi-attn + knowledge	CRF	T-2015, T-2017
ITA [19]	2022	BERT-base	ViT	Image-text prefix align.	CRF	T-2015, T-2017
MLNet [20]	2023	BERT-base	ResNet	Multi-level filter gate	CRF	T-2015, T-2017
MGCMT [5]	2024	BERT-base	ResNet + Mask RCNN	Multi-granularity CM Transformer	CRF	T-2015, T-2017
MAHE [21]	2025	BERT-base	ResNet	Multiscale hybrid expert	CRF	T-2015, T-2017
TriMod Fusion [22]	2025	BERT-base	ViT	Three-modal fusion	CRF	T-2015, T-2017
MSCMT Hybrid (Ours)	**2025**	**BERTweet-base**	**CLIP ViT-L/14**	**Token-level sigmoid gating**	**CRF**	**T-2015 + T-2017 (merged)**

G. Classifier and CRF Decoder

A linear classifier maps each fused representation to emission scores:

$$e_i = W_c f_i^{\text{fused}} + b_c, \quad W_c \in \mathbb{R}^{9\times 1536}, \quad e_i \in \mathbb{R}^9$$

A Conditional Random Field (CRF) [23] models the joint probability of the complete label sequence:

$$p(Y|E;\theta_{\text{CRF}}) \propto \exp\left(\sum_i [e_i(y_i) + A(y_{i-1}, y_i)]\right)$$

where $A \in \mathbb{R}^{9\times 9}$ is a learned transition matrix encoding label dependencies and enforcing BIO constraints (e.g., penalising `I-PER` following `O`).

Training minimises the negative log-likelihood of the correct label sequence. During inference, the Viterbi algorithm decodes the globally optimal sequence in $\mathcal{O}(M \cdot |\mathcal{Z}|^2)$ time.

IV. EXPERIMENTS AND RESULTS

A. Datasets and Class Balancing

We evaluate on a merged corpus formed by combining TWITTER-2015 [2] and TWITTER-2017 [3]. TWITTER-2015 contains 8,257 tweets with 12,800 entity annotations, while TWITTER-2017 provides 4,819 tweets with 8,724 annotations. Both datasets use the same four-class schema: `PER`, `LOC`, `ORG`, and `OTHER`. Merging the two corpora yields approximately 8,890 training instances and 3,200 test instances, enlarging the available training signal and exposing the model to a more diverse distribution of entities and image types. The merged training set exhibits significant class imbalance, with `O` tokens dominating and `ORG` and `OTHER` underrepresented. We address this using PyTorch's `WeightedRandomSampler`, which samples training instances with probability inversely proportional to class frequency, ensuring that all entity types receive adequate gradient signal across epochs. Table II summarises the sentence and entity counts across both datasets and the merged corpus. The imbalance between entity classes - most visibly between PER and OTHER - directly motivates the class-weighted sampling strategy described below.

B. Hyperparameter Optimisation

Following the recommendations of Devlin et al. [13] for fine-tuning BERT-family models, we searched learning rates in the range [1e-5, 5e-5] and batch sizes $\{8, 16, 32\}$. All configurations were trained for 20 epochs using AdamW [24], with weighted F1 on a held-out validation split as the model selection criterion.

The 20-epoch limit was determined empirically by monitoring validation performance, which plateaued and subsequently degraded beyond this point - indicating overfitting consistent with the relatively small dataset size. Table III summarises the grid search results.

The results reveal a clear preference for a moderate batch size (16) and a conservative learning rate (3e-5). Larger batches reduce gradient update frequency below what is required for stable convergence on this dataset scale, while smaller batches introduce excessive gradient noise. The winning configuration (batch size 16, learning rate 3e-5) is used in all subsequent experiments.

C. Evaluation Protocol

All experiments are evaluated using the span-level CoNLL evaluation scheme, implemented via the seqeval library. Under this protocol, a predicted entity span is counted as correct only when both its boundary and its type match the gold annotation exactly. A span with the correct boundary but the wrong type receives no credit; a boundary error of even one token simultaneously generates a false positive and a false negative. This is the same strict evaluation protocol used by every baseline reported in Table I, ensuring that our results are directly comparable at the metric level. We report precision, recall, and F1 at the per-entity level (PER, LOC, ORG, OTHER) as well as micro-averaged and weighted-averaged F1 across all entity types. Weighted F1 is the primary comparison

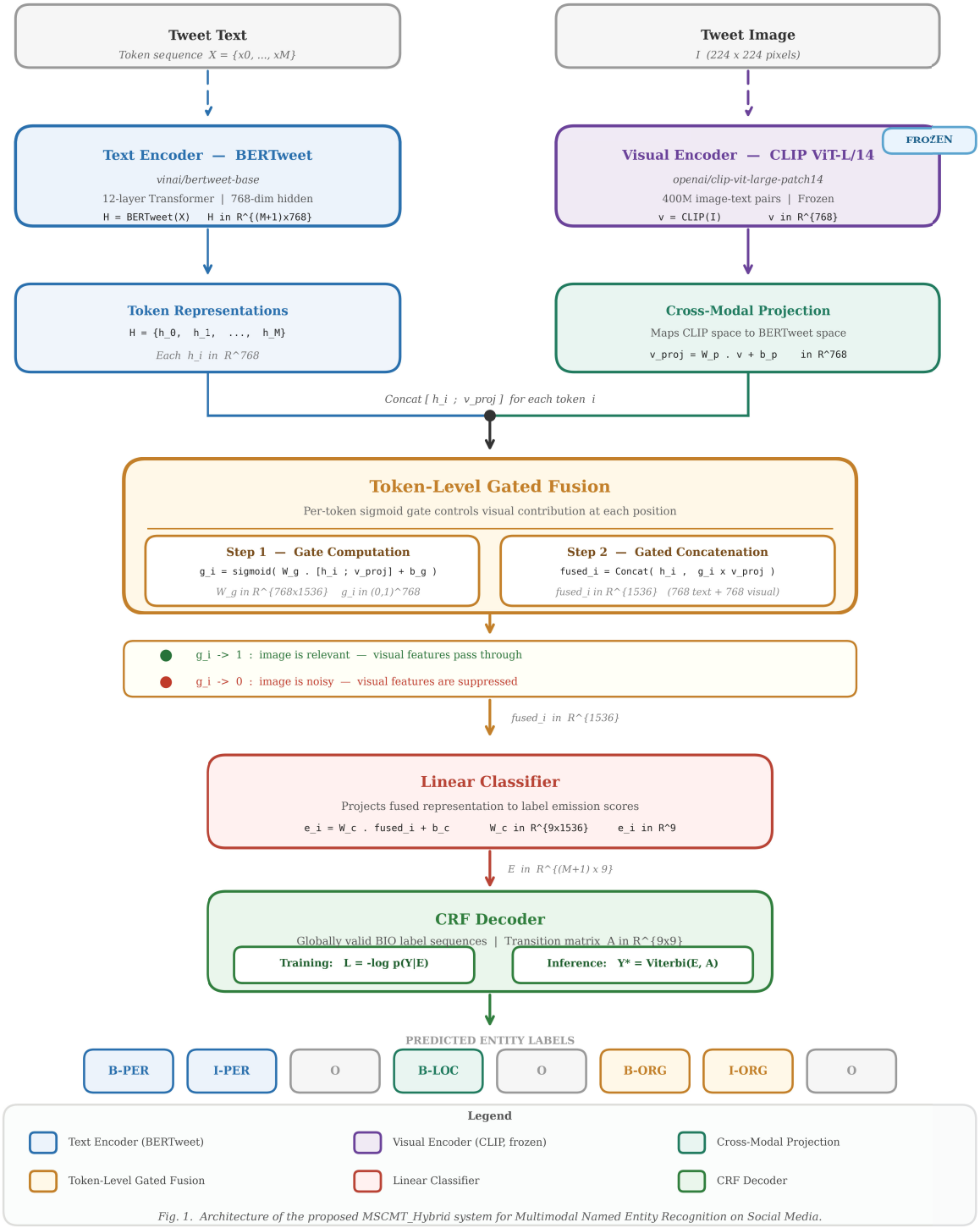

Fig. 1. Architecture of the proposed MSCMT_Hybrid system for Multimodal Named Entity Recognition on Social Media.

Fig. 1. Architecture of the proposed MSCMT Hybrid system for Multimodal Named Entity Recognition on Social Media.

metric, as it accounts for the support imbalance between entity classes in the merged corpus.

D. Main Results

Table IV reports per-entity performance of MSCMT_Hybrid on the merged test set.

Performance is strongest on PER (F1 = 0.8619) and LOC (F1 = 0.8258), both of which are visually grounded entity types. Person faces and geographical landmarks frequently appear in tweet images, providing reliable discriminative signal for the gating mechanism.

TABLE II. STATISTICS OF THE TWITTER-2015, TWITTER-2017, AND MERGED DATASETS

Dataset	Split	Sent.	PER	LOC	ORG	OTHER	Total
TWITTER-2015	Train	4,000	2,217	2,091	928	1,364	6,600
	Dev	1,000	552	521	247	305	1,625
	Test	3,257	1,816	1,697	839	885	5,237
TWITTER-2017	Train	3,373	2,546	1,858	1,193	948	6,545
	Dev	723	534	403	248	204	1,389
	Test	723	519	409	247	186	1,361
Merged	**Train**	**8,890**	**5,224**	**4,280**	**2,312**	**2,074**	**13,890**
	Test	**3,200**	**1,873**	**1,720**	**860**	**762**	**5,215**

Note: Sent. = Sentences. PER = Person, LOC = Location, ORG = Organisation, OTHER = Miscellaneous. Entity counts reflect BIO-tagged mentions in the merged training and test sets.

TABLE III. GRID SEARCH RESULTS ON MERGED TEST SET

Configuration	Prec.	Rec.	F1 (%)	Best
Batch=16, LR=3e-5	0.7368	0.7778	75.97	✓
Batch=16, LR=5e-5	0.7433	0.7728	75.43	
Batch=16, LR=3e-4	0.7368	0.7778	75.59	
Batch=32, LR=3e-5	0.7343	0.7540	74.63	
Batch=8, LR=3e-5	0.7281	0.7873	73.57	

TABLE IV PER-ENTITY RESULTS OF MSCMT_HYBRID

Entity	Prec.	Rec.	F1	Support
PER	0.8328	0.8932	0.8619	1873
LOC	0.7968	0.8570	0.8258	1720
ORG	0.6418	0.6605	0.6510	860
OTHER	0.4965	0.4685	0.4821	762
Micro avg	0.7453	0.7728	0.7588	5215
Weighted avg	0.7368	0.7778	0.7597	5215

The `ORG` class (F1 = 0.6510) suffers from limited support (860 instances) and the abstract visual representation of organisations. Brand logos and corporate imagery are more ambiguous for a vision encoder than faces or physical landmarks. The `OTHER` class (F1 = 0.4821) is inherently the most challenging due to its semantic heterogeneity.

E. Ablation Study

The strong performance of MSCMT_Hybrid raises a natural question: which components are actually doing the work? To answer this, we isolate the two most consequential design choices—the domain-specific text encoder and the token-level gated fusion mechanism—and test degraded variants against the merged Twitter-2015 + Twitter-2017 benchmark.

Throughout all conditions, CLIP ViT-L/14 is kept *frozen*. CLIP was pre-trained on 400 million image-text pairs; fine-tuning it on the ≈8,890 samples available here would cause catastrophic forgetting—the well-documented failure mode in which a large pre-trained model, exposed to a small target dataset, overwrites broadly useful representations with narrow, dataset-specific ones that do not generalise. Freezing CLIP ensures that any performance difference between variants reflects only the text encoder or the fusion strategy.

Variant 1: $\text{BERT}_{\text{base}}$ + Simple Fusion. Both proposed contributions are removed simultaneously. BERTweet is replaced with `bert-base-uncased`, and the sigmoid gate is replaced with direct element-wise addition of the projected visual embedding and each token representation. This variant is the strongest possible internal baseline: it tells us what the architecture achieves before either contribution is introduced.

Variant 2: $\text{BERT}_{\text{base}}$ + Gated Fusion. The sigmoid gate is restored, but the text encoder remains $\text{BERT}_{\text{base}}$. This isolates the fusion mechanism's contribution independently of pre-training domain. If the gate is genuinely useful regardless of the encoder, it should produce gains over Variant 1 even without BERTweet.

Table V reports the results. The full model outperforms Variant 1 by **3.67 percentage points** in weighted F_1 (75.97% vs. 72.30%). The more instructive comparison, however, is between Variants 1 and 2: restoring the gate while keeping $\text{BERT}_{\text{base}}$ produces a slight *drop* to 71.83%—the opposite of what one would expect if the gate were unconditionally helpful.

This is a diagnostic result, not a flaw. The gate conditions its per-token decisions jointly on the token's linguistic representation and the projected visual embedding. $\text{BERT}_{\text{base}}$, pre-trained on Wikipedia and BooksCorpus, systematically fragments entity-bearing tokens such as hashtags, handles, and misspelled proper nouns into uninformative subword pieces—degraded representations at exactly the positions where the gate needs reliable signal. Faced with noisy token encodings, the gate defaults to inconsistent behaviour that suppresses visual information where it would genuinely help, explaining the marginal regression from Variant 1 to Variant 2.

BERTweet changes this entirely. Pre-trained on 850 million English tweets, it produces stable, semantically meaningful representations for the exact surface forms that $\text{BERT}_{\text{base}}$ cannot handle. With these as input, the gate receives strong linguistic grounding for its per-token decisions, and the full model achieves the observed 3.67-point gain. The two compo-

nents are mutually enabling rather than independently additive: the gate's selectivity is useful only when the text encoder is expressive enough to drive it, and BERTweet reaches its full potential only when paired with a fusion strategy that incorporates visual evidence selectively rather than uniformly. Removing either degrades performance; combining both is what yields the result.

TABLE V. ABLATION RESULTS ON THE MERGED TWITTER-2015 + TWITTER-2017 TEST SET. CLIP IS FROZEN THROUGHOUT. **BOLD** = BEST PER COLUMN.

Variant	Type	P	R	F_1	Wt. F_1
V1: $BERT_{base}$ + Simple Fusion	PER	0.8079	0.8868	0.8455	72.30%
	LOC	0.7813	0.8349	0.8072	
	ORG	0.6379	0.5837	0.6096	
	OTHER	0.3356	0.3871	0.3595	
V2: $BERT_{base}$ + Gated Fusion	PER	0.8234	0.8665	0.8444	71.83%
	LOC	0.7696	0.8291	0.7982	
	ORG	0.5970	0.5977	0.5973	
	OTHER	0.3463	0.3845	0.3644	
MSCMT_Hybrid (BERTweet + Gated)	PER	**0.8328**	**0.8932**	**0.8619**	**75.97%**
	LOC	**0.7968**	**0.8570**	**0.8258**	
	ORG	**0.6418**	**0.6605**	**0.6510**	
	OTHER	**0.4965**	0.4685	**0.4821**	

F. Per-Dataset Generalisation

To contextualise the merging decision and assess how well the model generalises across the two corpora, we conduct a cross-dataset evaluation. In the first configuration, the model is trained exclusively on TWITTER-2015 and evaluated on the TWITTER-2017 test set; in the second, training is on TWITTER-2017 and evaluation on TWITTER-2015. The results are reported in Table VI alongside the merged-corpus result from Section IV-C.

Training on TWITTER-2015 alone and testing on TWITTER-2017 yields a average F1 score of 63.96%. The drop is most pronounced on ORG, where recall falls to 0.37 despite reasonable precision—a pattern consistent with the entity distribution mismatch between the two corpora: TWITTER-2017 carries a substantially larger proportion of organisation mentions (8,724 total annotations vs. 6,600 in TWITTER-2015), and a model that has not encountered this diversity of organisation references during training predictably struggles to recall them at test time. PER transfers reasonably well across datasets (F1 = 0.8578), which is unsurprising given that person-face imagery is a more corpus-agnostic visual cue than brand logos or organisational symbols. The reverse configuration—training on TWITTER-2017 and testing on TWITTER-2015—yields an average F1 of 64.17%, a comparably weak result. Here the asymmetry shifts: LOC recall improves substantially (0.8356) while ORG precision degrades to 0.4942. TWITTER-2015 contains a higher density of location-oriented posts, and the CLIP visual encoder's landmark representations generalise well in this direction, but the precision loss on ORG reflects the noisier visual grounding of organisational entities when training signal is drawn from a different distribution. Across both conditions, the cross-dataset average F1 stays in the 63–64% range.. The merged-corpus model improves this to 75.97%-a gain of roughly 12 percentage points-because training on the combined data exposes the model to the full entity type distribution it will encounter at test time. The ORG gap is the clearest illustration: pooling both corpora more than doubles the ORG training signal, and the gating mechanism can learn more reliable visual cues for organisational entities only when sufficient variety is available during training. These results confirm that the merging decision is not a matter of convenience but a practical necessity given the complementary coverage of the two datasets.

G. Discussion

MSCMT_Hybrid achieves a weighted F1 of 75.97% on the merged TWITTER-2015 and TWITTER-2017 benchmarks, with particularly strong performance on person and location entities where visual context provides consistent disambiguating cues. A central finding of this work is that encoder quality matters more than fusion complexity. While several existing MNER systems rely on multi-scale object detection, graph-based entity linking, or multi-layer cross-modal Transformers, MSCMT_Hybrid deliberately employs two well-chosen pre-trained encoders combined with a lightweight gating mechanism. The ablation study in Section IV-G corroborates this directly: replacing BERTweet with BERT-base costs 3.67 weighted F1 points regardless of whether the gate is present, whereas the gate alone-without a domain-matched text encoder, produces no gain over simple fusion. Encoder quality is therefore the primary driver, and fusion granularity amplifies it rather than substituting for it. The per-dataset evaluation in Section IV-F adds a further dimension to this picture. Training on either corpus in isolation and testing on the other yields weighted F1 scores in the 63–64% range, compared to 75.97% on the merged corpus. The gap is largest on ORG, the entity class with the most skewed distribution across the two datasets, confirming that the merged training setup is not merely a convenience but a practical necessity for covering the full entity type distribution encountered at test time. Taken together, these results suggest that domain-specific textual pre-training paired with a large-scale visually trained encoder contributes more substantially to performance than architectural sophistication in the fusion module. This finding has practical value, as simpler architectures are easier to reproduce, deploy, and extend.

H. Analysis of VLM Adaptation Failure

The results reported above were obtained with the hybrid design. We now examine why a more direct approach-adapting unified vision-language models end-to-end-proves inadequate for this task. Prior to developing the hybrid architecture, we investigated whether end-to-end vision-language models (VLMs) could serve as a direct foundation for MNER. Two

TABLE VI. CROSS-DATASET AND MERGED-CORPUS EVALUATION OF MSCMT HYBRID

Train	Test	Entity	Prec.	Rec.	F1
T-2015	T-2017	PER	0.9256	0.7993	0.8578
T-2015	T-2017	LOC	0.7468	0.6978	0.7215
T-2015	T-2017	ORG	0.7398	0.3699	0.4932
T-2015	T-2017	Micro avg	0.6576	0.5740	0.6130
T-2015	T-2017	Weighted avg	0.7444	0.5740	0.6396
T-2017	T-2015	PER	0.7858	0.8687	0.8252
T-2017	T-2015	LOC	0.7220	0.8356	0.7746
T-2017	T-2015	ORG	0.4942	0.6660	0.5674
T-2017	T-2015	Micro avg	0.6435	0.6926	0.6672
T-2017	T-2015	Weighted avg	0.5988	0.6926	0.6417
T-2015+T-2017	Test	Weighted avg	**0.7368**	**0.7778**	**0.7597**

experimental pipelines were constructed: one using CLIP ViT-L/14 as both visual and textual encoder, and another based on BLIP-2, following standard fine-tuning protocols. `WeightedRandomSampler` was applied in both configurations to address class imbalance.

Both VLM-based configurations stagnated at test accuracies of approximately 49–52%, despite normal convergence of training loss. We attribute this to three primary factors.

First, CLIP and BLIP-2 generate a single global `[CLS]` representation per image, lacking positional decomposition aligned with token positions. NER requires token-level predictions, and a shared global vector provides no positional specificity.

Second, the pre-training objectives of these models (contrastive alignment and generative captioning) bear no structural resemblance to BIO sequence labeling. Their internal representations are therefore not optimized for boundary detection.

Third, injecting global representations into a sequence labeling framework leads to a degenerate optimum: predicting the majority `O` class for most tokens. This minimizes cross-entropy loss but results in poor F1 performance on minority entity classes.

MSCMT_Hybrid avoids these failure modes by freezing CLIP to preserve its visual semantics, injecting visual information at each token position through gated fusion, and leveraging BERTweet representations shaped by token-level masked language modeling objectives that are inherently compatible with NER.

V. CONCLUSION

This paper presented MSCMT_Hybrid, a hybrid architecture for multimodal named entity recognition (MNER) on social media. By pairing BERTweet with a frozen CLIP ViT-L/14 encoder and bridging the two modalities through a lightweight token-level sigmoid gate, the model achieves a weighted F1 of 75.97% on the merged TWITTER-2015 and TWITTER-2017 benchmark; a result competitive with published state-of-the-art systems, obtained with a substantially simpler fusion design. Three broader findings emerged from this work. First, domain alignment of the text encoder is critical. Replacing BERT-base with BERTweet—pre-trained on 850 million tweets-aligns the language model's prior distribution with social media text, improving entity recognition without increasing architectural complexity. Second, the quality and scale of visual pre-training carry more impact than fusion sophistication. CLIP ViT-L/14, trained on 400 million image-text pairs, produces sufficiently rich visual representations such that a single linear projection and sigmoid gate are adequate to exploit cross-modal signal effectively. The ablation study further showed that these two components are mutually enabling: the gate's selectivity is meaningful only when the text encoder produces representations expressive enough to drive it, and neither contributes independently what both achieve together. Third, the per-dataset evaluation demonstrated that training on either corpus in isolation leaves a systematic gap on underrepresented entity classes—most visibly ORG-that only the merged training setup resolves, providing an empirical justification for the corpus combination strategy rather than a post-hoc one. We also documented a systematic failure mode that arises when unified vision-language models, designed primarily for global sentence-level tasks, are directly adapted to token-level structured prediction. This negative result serves as a practical contribution, offering a replicable reference point and a principled explanation for why hybrid architectures remain necessary for MNER, even as general-purpose VLMs continue to advance. Future work includes exploring cross-attention between token representations and CLIP patch-level features to enable spatially resolved visual grounding, applying the architecture to low-resource languages where social media NER data is scarce, and extending the gating mechanism to selectively integrate multiple image regions aligned to different entity spans within a single tweet.

ACKNOWLEDGMENT

The authors acknowledge funding support from the UK Research and Innovation (UKRI) through Project APP47457, titled "Super-efficient Sustainable Cooling Solution for All Applications (S2Cool)", under the Ayrton Challenge Programme.

REFERENCES

[1] S. Moon, L. Neves, and V. Carvalho, "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), 2018, pp. 852–860.

[2] 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.

[3] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[4] J. Yu, J. Jiang, L. Yang, and R. Xia, "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*, 2020, pp. 3342–3352.

[5] P. Liu, G. Wang, H. Li, J. Liu, Y. Ren, H. Zhu, and L. Sun, "Multi-granularity cross-modal representation learning for named entity recognition on social media," *Information Processing & Management*, vol. 61, no. 1, p. 103546, 2024.

[6] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 9–14. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.2/

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[8] J. Li, D. Li, S. Savarese, and S. Hoi, "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*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 19 730–19 742. [Online]. Available: https://proceedings.mlr.press/v202/li23q.html

[9] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," 2015. [Online]. Available: https://arxiv.org/abs/1508.01991

[10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Knight, A. Nenkova, and O. Rambow, Eds. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 260–270. [Online]. Available: https://aclanthology.org/N16-1030/

[11] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional

[16] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados, "Timelms: Diachronic language models from twitter," in *Proceedings of the 60th Annual Meeting of the Association for Computational* lstm-cnns-crf," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1064–1074.

[12] M. E. Peters *et al.*, "Deep contextualized word representations," in *NAACL-HLT*, 2018.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[14] Z. Wu, C. Zheng, Y. Cai, J. Chen, H.-f. Leung, and Q. Li, "Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts," in *Proceedings of the 28th ACM International conference on multimedia*, 2020, pp. 1038–1046.

[15] J. Li and F. Fukumoto, "Multi-task neural shared structure search: A study based on text mining," in *International Conference on Database Systems for Advanced Applications*. Springer, 2021, pp. 202–218. *Linguistics: System Demonstrations (ACL 2022)*, Dublin, Ireland, May 2022, pp. 251–260.

[17] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Liu, C. Huang, B. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, New Orleans, LA, USA, Jun. 2022, pp. 18 082–18 091.

[18] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *ECCV*, 2022, pp. 696–712.

[19] X. Wang, M. Gui, Y. Jiang, Z. Jia, N. Bach, T. Wang, Z. Huang, and K. Tu, "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 (NAACL-HLT 2022)*, Seattle, WA, USA, Jul. 2022, pp. 3176–3189.

[20] H. Zhai, X. Lv, Z. Hou, X. Tong, and F. Bu, "Mlnet: A multi-level multimodal named entity recognition architecture," *Frontiers in Neurorobotics*, vol. 17, p. 1181143, Jun. 2023.

[21] S. Yang, B. Mo, D. Liu, and L. Zhu, "Mahe: a multiscale and hybrid expert-based model for image-text enhanced named entity recognition on social media," *Scientific Reports*, vol. 15, no. 1, p. 17663, 2025.

[22] M. Alfaqeeh, "Trimod fusion for multimodal named entity recognition in social media," in *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*. IEEE, 2024, pp. 1–9.

[23] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning (ICML)*, 2001, pp. 282–289.

[24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.