ON EVALUATING THE ADVERSARIAL ROBUSTNESS OF FOUNDATION MODELS FOR MULTIMODAL ENTITY LINKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal entity linking (MEL) aims to identify a target entity from a set of semantically related candidates using visual and textual evidence. Unlike conventional image classification, MEL requires fine-grained candidate disambiguation, and small changes in cross-modal similarity may reverse the ranking between the correct entity and hard negatives. However, the adversarial robustness of MEL remains underexplored. We conduct a systematic evaluation of visual adversarial robustness under two candidate-constrained settings: Image-to-Text Entity Linking (I2T-EL) and Image+Text-to-Text Entity Linking (IT2T-EL). Our evaluation covers general-purpose vision-language encoders, multimodal large language models, three adversarial attacks, and five MEL benchmarks. The results show that visual perturbations can substantially reduce linking accuracy, although the degradation varies across models, datasets, and attack strengths. We further observe that informative textual context can partially preserve entity-level semantics under visual attacks. Motivated by this observation, we propose LLM-RetLink, a training-free framework that converts visual evidence into multiple discrete entity hypotheses, retrieves candidate-grounded knowledge, filters the retrieved evidence, and performs evidence-aware entity linking. Experiments show that LLM-RetLink improves both clean and adversarial linking accuracy by 0.4%-35.7% across the evaluated settings. Our study provides an adversarial MEL benchmark and highlights a new direction for improving entity-level robustness through discrete semantic grounding.

1 INTRODUCTION

With the explosive growth of multimodal information on the Internet, Multimodal Entity Linking has emerged as a crucial bridge connecting visual content with structured knowledge. It plays a central role in cross-modal information understanding and has been widely applied in various downstream tasks such as knowledge-enhanced question & answering Hu et al. (2024), image-text retrieval Huang et al. (2024), and open-domain entity alignment Qiu et al. (2024). In recent years, driven by advances in pretrained large language models, large vision models, and retrieval-augmented mechanisms, MEL has made significant progress in cross-modal representation learning, candidate generation, and entity disambiguation strategies Pourreza et al. (2024); Pons et al. (2024). These developments have enabled MEL to consistently achieve state-of-the-art performance on multiple standard benchmark datasets, gradually establishing MEL as a key paradigm for achieving general image-text semantic understanding.

Despite recent advances in MEL, the impact of adversarial examples remains underexplored, particularly in terms of systematic evaluation against visual adversarial perturbations. Most existing MEL methods assume that visual inputs are clean and accurate, while in real-world scenarios, images are often subject to noise or even malicious attacks Chi et al. (2024). Meanwhile, adversarial attacks have demonstrated strong deceptive capabilities in other multimodal tasks such as image classification and visual question answering Archana & Jeevaraj (2024); Sima et al. (2024). However, there is still a lack of systematic investigation of the vulnerability of MEL models under such perturbations. This gap hinders the reliable deployment of MEL models in complex real-world environments.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

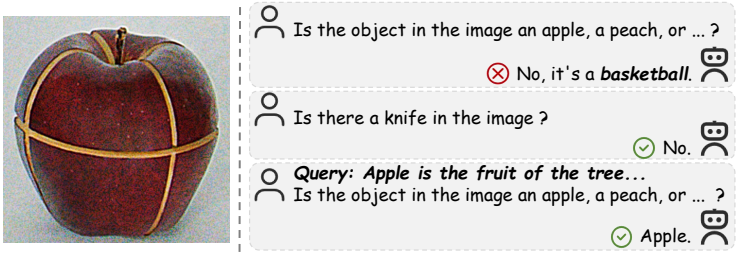


Figure 1: QA pairs for LLaVA given an adversarial image. In the first two QA pairs, LLaVA provides entirely incorrect answers. However, it correctly responds to subsequent questions that are unrelated to the attacked object (apple). Notably, the contrast between the first and last QA pairs highlights that LLaVA can answers the question correctly once additional context is introduced. These observations support key findings in this paper. Source: M^3EL Wang et al. (2025).

MEL differs from conventional visual classification and open-ended visual question answering in an important aspect: the model must distinguish the correct entity from a set of semantically related candidates. Therefore, an adversarial perturbation does not need to completely change the perceived visual category. It may cause a linking error simply by reducing the score margin between the correct entity and a hard negative. For example, a perturbation may shift the prediction from "apple fruit" to another visually similar fruit, or from the fruit entity "Apple" to a same-name entity such as Apple Inc. This candidate-ranking nature raises three questions. 1) How do visual perturbations affect the ranking margin between the correct entity and hard negatives? 2) Under what conditions can textual context compensate for corrupted visual evidence? 3) Can a training-free semantic grounding mechanism improve robustness without directly modifying the underlying vision-language model? Existing studies have not systematically answered these questions in the context of MEL.

To answer these questions, we establish a unified evaluation protocol covering two candidate-constrained MEL settings, I2T-EL and IT2T-EL. We evaluate general-purpose vision-language encoders adapted to candidate ranking, together with multimodal large language models, on five public MEL datasets. Rather than only measuring accuracy degradation, we compare model behavior across attack types, attack strengths, datasets, and input modalities. The results show that robustness varies substantially across architectures and datasets. Nevertheless, most evaluated configurations suffer clear performance degradation under sufficiently strong attacks. We also find that informative textual context can reduce, but does not eliminate, the effect of visual perturbations.

Overall, our main contributions are as follows:

- We formulate visual adversarial robustness in MEL as a candidate-ranking problem and establish a unified evaluation protocol for I2T-EL and IT2T-EL across five public datasets.
- We evaluate both general-purpose vision-language encoders adapted to MEL and multimodal large language models, revealing how attack type, attack strength, textual context, and candidate ambiguity affect linking robustness.
- We construct adversarial variants of five MEL benchmarks and provide the generated examples and evaluation protocol to facilitate reproducible robustness research.
- We propose LLM-RetLink, a training-free semantic grounding framework that generates multiple discrete entity hypotheses, retrieves candidate-grounded evidence, filters unreliable evidence, and performs evidence-aware entity linking.

2 Related Work

Multi-modal Entity Linking The pioneering study by Moon Moon et al. (2018)first demonstrated the effectiveness of visual information in disambiguating short-text mentions on social networks, proposing a zero-shot learning framework that integrates textual, visual, and lexical features. Subsequently, Adjali Adjali et al. (2020)introduced an automated MEL dataset construction method based on Twitter data. Gan Gan et al. (2021) formulated the alignment between textual and visual mentions as a bipartite graph matching problem. Wang Wang et al. (2022) used a multi-modal co-attention mechanism to hierarchically extract jointly attended textual and visual features.

Zhang Zhang & Huang (2022) proposed an IMLN to make full use of information in terms of mentions and knowledge. Xing Xing et al. (2023b) introduced DRIN to capture fine-grained relationships between entities and mentions. Luo Luo et al. (2023a) proposed the MIMIC, which is designed to comprehensively model both abbreviated textual context and implicit visual cues, enabling more robust multimodal representation learning. Long Long et al. (2024) introduced GELR to incorporate a knowledge retriever to enhance the visual information of the entity. Song Song et al. (2024) and Zhao Zhao et al. (2025) implemented queries with multimodal data and addressed semantic gaps using cross-modal enhancers between text and image information.

Adversarial attacks Adversarial attacks refer to the addition of imperceptible perturbations to input, which mislead neural networks into making incorrect predictions. Based on the amount of information accessible to the attacker, these attacks can be categorized into three types: white-box attacks Szegedy et al. (2014); Carlini & Wagner (2017); Athalye et al. (2018), where the attacker has full access to the model architecture and parameters; black-box attacks Papernot et al. (2017); Su et al. (2017), where only input-output queries are available and internal details remain hidden; and gray-box attacks, where the attacker has partial knowledge of the model, such as its architecture.

Adversarial Examples Neural network models have been shown to be highly vulnerable to adversarial examples in Computer Vision tasks Carlini & Wagner (2017); Madry et al. (2018); Croce & Hein (2020). These adversarial examples are typically crafted by applying subtle and carefully designed perturbations to input data, which remain visually indistinguishable from the original but are sufficient to cause incorrect predictions. We evaluated the effectiveness of several well-known adversarial attack methods in manipulating model outputs, including Projected Gradient Descent (PGD) Madry et al. (2018), Auto-PGD (APGD) Croce & Hein (2020), and the Carlini & Wagner (CW) attacks Carlini & Wagner (2017). Studying these attacks deepens understanding of model robustness and guides the development of defenses.

3 METHOD

3.1 PROBLEM FORMULATION

We consider gradient-based white-box adversarial attacks. Given an input-label pair (x, y) and a classifier f, the objective is to find a perturbation δ such that the prediction of the perturbed input is incorrect:

$$f(x + \delta) \neq y \tag{1}$$

PGD seeks to maximize the classification loss while constraining the perturbation within an L_∞ ball:

$$\max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}(f(x + \delta), y) \tag{2}$$

where $\mathcal{L}$ denotes the cross-entropy loss, and ϵ controls the attack strength.

CW attack formulates the problem as follows:

$$\min_{\delta} \|\delta\|_p + c \cdot g(x + \delta), \quad x + \delta \in [0, 1]^n \tag{3}$$

where $c > 0$ is a balancing hyperparameter and the function g is defined as:

$$g(x + \delta) = \max\left(f(x + \delta)_y - \max_{i \neq y} f(x + \delta)_i, -\kappa\right) \tag{4}$$

where, κ is the confidence margin controlling the strength of misclassification.

3.2 ATTACKS & ADVERSARIAL GENERATION

We adopt PGD and CW as two representative gradient-based attacks, and include APGD as a variant of PGD. For each method, we consider two configurations, normal and strong, distinguished by the perceptual visibility of the perturbations. Detailed parameters are summarized in Table 1.

Normal. We follow previous works Cui et al. (2024), setting the constraint for CW to 20 and the $\epsilon = 8/255$ for PGD / APGD. For brevity, we denote this attack as N.

Table 1: Parameters for the attacks under Normal and Strong settings. Dist. refers to distance measure, c and κ refers to the constraint and confidence in Cui et al. (2024).

Type	Method	Step	Size	ϵ	Dist.	c	κ
Normal	PGD	20	2/255	8/255	L_∞	-	-
	APGD	20	-	8/255	L_∞	-	-
	CW	50	0.01	-	L_2	20	0
Strong	PGD	40	2/255	0.2	L_∞	-	-
	APGD	40	-	0.2	L_∞	-	-
	CW	75	0.05	-	L_2	100	0

Strong. We set the constraint for CW to 100 and the $\epsilon = 0.2$ for PGD and APGD. We term this attack S.

Generation. We generate adversarial samples on five datasets: Wikidata-MEL, Richpedia-MEL, WikiDiverse, WIKIPerson, and M^3EL. For PGD and APGD, we maximize the Cross-Entropy loss between the model logits and the ground-truth label. For CW, we minimize the sum of the l_2 distance of the perturbation δ and the f-function from the original paper Cui et al. (2024). Figure **??** shows adversarial samples.

As shown in Figure 2, we first used the text encoder to encode the text entity labels in the format of "*a photo of <entity>*". Then, we computed the cosine similarity between the image encoding and encoded labels and used the result as the prediction logit for adversarial generation and evaluation.

3.3 Tasks & Evaluation

To evaluate the impact of visual adversarial on MEL models, we consider two commonly used tasks: Image-to-Text Entity Linking and Image+Text-to-Text Entity Linking. The detailed evaluation process for the specific task is as follows.

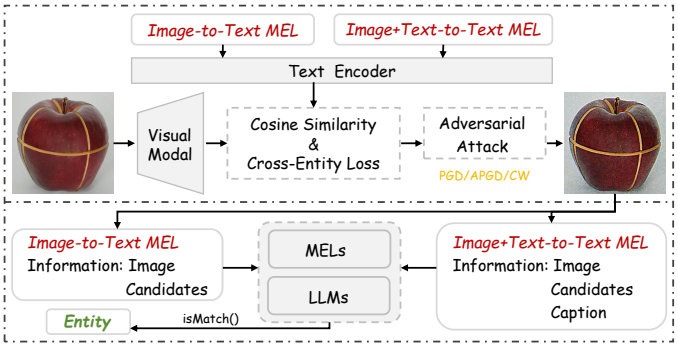


Figure 2: Overview of our procedure for attack generation and evaluation over I2T MEL and IT2T MEL.

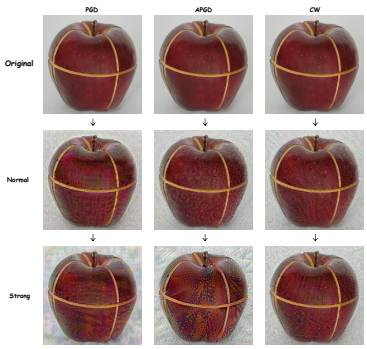


Figure 3: The samples of adversarial images.

3.3.1 I2T-EL: Image-to-Text Entity Linking

We systematically evaluate the robustness of I2T-EL under visual adversarial perturbations using images and corresponding candidates from five standard datasets. The evaluation covers both traditional MEL models and state-of-the-art Multimodal Large Language Models (MLLMs).

For traditional MEL methods, we first construct all candidates in the format of "*a photo of <candidate >*", and obtain their embeddings via a text encoder. We then compute the cosine similarity between the image embedding and each candidate embedding, and select the entity with the highest similarity score as the model prediction. The final linking accuracy is evaluated based on this prediction.

For MLLMs, the adversarial image and candidates are jointly provided as input, along with a tailored prompt: "*Which candidate best describes the main object in this image? \n Answer with the index of the candidate.*". The linking performance is then assessed based on the model's response.

3.3.2 IT2T-EL:IMAGE+TEXT-TO-TEXT ENTITY LINKING

The robustness of IT2T-EL is evaluated on five standard datasets, using images, entity descriptions, and candidates, following the same baselines as in the I2T-EL task.

For traditional MEL methods, we first obtain the embeddings of the candidates using the same procedure. The entity description is encoded using the same text encoder and fused with the image embedding through element-wise addition to form a unified multimodal representation. We then compute the cosine similarity between this fused representation and each candidate embedding, selecting the candidate with the highest similarity as the prediction to assess linking accuracy.

For MLLMs, the adversarial image, entity description, and candidates are jointly used to construct the prompt: "*Please observe the image and understand the description. Which candidate best describes the main object in the image? \n Answer with the index of the candidate.*".

3.4 LLM AND RETRIEVAL-AUGMENTED ENTITY LINKING

We propose **LLM-RetLink**, a training-free framework for robust candidate-constrained MEL, as illustrated in Figure 4. Its central idea is to reduce the direct dependence of entity linking on a fragile continuous visual representation. Instead of using a single visual embedding as the final decision signal, LLM-RetLink first converts the image into multiple discrete entity hypotheses. It then collects candidate-grounded evidence, filters evidence inconsistent with the image and textual context, and performs the final linking decision.

Given an image I, textual context T, and candidate set $\mathcal{C}$, the framework contains four stages: entity hypothesis generation, candidate-grounded retrieval, evidence reliability filtering, and evidence-aware entity linking. No model parameter is updated during inference.

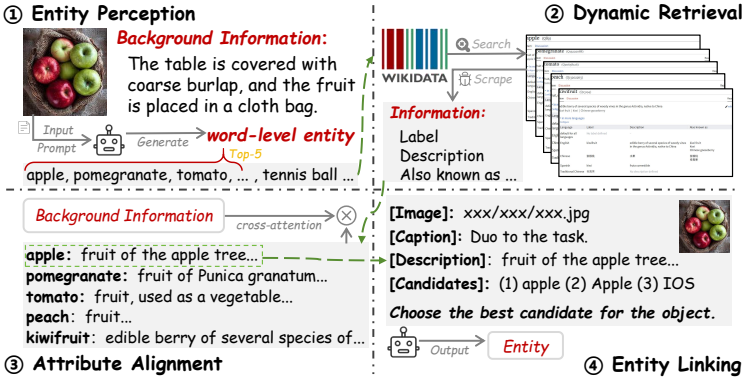


Figure 4: The overview of LLM-RetLink, which comprises four modules: Entity Perception, Dynamic Retrieval, Attribute Alignment and Entity Linking.

3.4.1 ENTITY HYPOTHESIS GENERATION

We use an MLLM to generate a background description of the image and a ranked list of k possible entity hypotheses:

$$H(I) = \{(h_j, p_j)\}_{j=1}^{k}, \tag{5}$$

where h_j is a discrete entity hypothesis and p_j is the model-reported confidence or normalized ranking score. We retain multiple hypotheses because an adversarial image may cause the top-1 visual prediction to be incorrect, while the correct entity may still appear among lower-ranked hypotheses.

The prompt requires the model to describe observable visual attributes before producing entity names, including object type, color, shape, scene, and associated context. We use $k = 5$ by default and evaluate the sensitivity to k in the experiments.

3.4.2 CANDIDATE-GROUNDED RETRIEVAL

Rather than performing unrestricted web search using only the generated visual hypothesis, we ground retrieval in the given candidate set. For each hypothesis h_j and candidate entity e_i, we construct a query using the hypothesis, candidate name, candidate type, and available textual context:

$$q_{ij} = [h_j; \mathrm{name}(e_i); \mathrm{type}(e_i); T]. \tag{6}$$

The retriever collects a short description, aliases, entity type, and salient attributes for each candidate from the same knowledge source. Using the same source and evidence fields for every candidate prevents the retriever from providing additional information only for a subset of candidates.

The retrieved evidence is denoted by $D_i = \{d_{i1}, \ldots, d_{ir}\}$. For reproducibility, we cache the retrieved documents, query strings, retrieval timestamps, and source identifiers used in the experiments.

3.4.3 EVIDENCE RELIABILITY FILTERING

External retrieval may return irrelevant, outdated, or conflicting information. We therefore score each retrieved evidence item before using it in final linking. For an evidence item d of candidate e_i, we compute

$$r(d, e_i) = \lambda_1 s_{\mathrm{ctx}}(d, I, T) + \lambda_2 s_{\mathrm{ent}}(d, e_i) + \lambda_3 s_{\mathrm{src}}(d) - \lambda_4 s_{\mathrm{conflict}}(d). \tag{7}$$

where s_{ctx} measures consistency with the image description and textual context, s_{ent} measures whether the evidence explicitly refers to candidate e_i, s_{src} represents source reliability, and s_{conflict} measures inconsistency with evidence retrieved from other sources.

We retain the top-ranked evidence items whose reliability scores exceed a threshold τ. If no evidence passes the threshold, the framework falls back to linking without external retrieval.

3.4.4 EVIDENCE-AWARE ENTITY LINKING

The final MLLM receives the original image, textual context, candidate entities, generated visual description, and filtered candidate-specific evidence. We use a structured prompt that requires the model to compare each candidate against the observed visual attributes and textual context.

The model is instructed to output only the selected candidate index and a confidence score. This constrained output format reduces parsing errors and avoids relying on unrestricted chain-of-thought generation. The final prediction is

$$\hat{e} = \mathrm{MLLM}(I, T, \mathcal{C}, H(I), \widetilde{D}), \tag{8}$$

where $\widetilde{D}$ denotes the filtered candidate evidence.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Datasets We conducted evaluations on three public English datasets, Wikidata-MEL Zhou et al. (2021), WikiDiverse Wang et al. (2022), and M^3EL Wang et al. (2025).

Baselines We categorize the evaluated models into 3 groups: general-purpose vision-language encoders (ALIGN Jia et al. (2021), BLIP Li et al. (2023), CLIP Radford et al. (2021), FLAVA Singh et al. (2022), OWL-ViT Minderer et al. (2022), and SigLIP Zhai et al. (2023)), which are adapted to MEL through candidate ranking rather than designed as end-to-end MEL systems; dedicated MEL models (DRIN Xing et al. (2023a), MIMIC Luo et al. (2023b), and GDMM Wang et al. (2023)), which are specifically developed for MEL and evaluated following their published preprocessing and candidate construction protocols; and based LLM linking models ChatEL Ding et al. (2024), GEMEL Shi et al. (2024), and LLMaEL Xin et al. (2024) and multimodal large language models (LLaVA Liu et al. (2023), Qwen2.5-VL Bai et al. (2025), and MiniGPT-4 Zhu et al. (2024)), which are tested using identical candidate sets and constrained output formats.

Implementation Details Experiments were run on A100 GPU (40GB), averaged over 3 runs. The temperature of MLLMS set to 0 and other parameters remaining at their default settings. We use Accuracy to evaluate MEL:

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(C_i \leftrightarrow E) \tag{9}$$

where M_i is the candidate, E is the correct entity, N is total instances, and the indicator function $\mathbb{I}(C_i \leftrightarrow E)$ denotes that C_i corresponds to entity E.

Table 2: Results on the robustness evaluation of different traditional MEL models across three datasets. Adversarial examples are used as input image along with contexts as input text. The numbers indicate the linking accuracy (%).

Dataset	Models	Image-to-Text							Image+Text-to-Text						
		RAW	PGD		APGD		CW		RAW	PGD		APGD		CW	
		P	N	S	N	S	N	S	P	N	S	N	S	N	S
Wikidata-MEL	ALIGN	30.1	26.8	22.3	28.5	27.2	24.2	21.8	28.3	27.4	27.1	27.5	27.0	27.3	26.9
	ALIGN*	52.2	50.7	49.4	50.7	47.3	51.1	37.6	47.3	46.6	46.3	46.3	46.1	46.7	44.1
	BLIP	45.4	44.5	42.4	43.6	42.8	43.3	42.7	67.3	45.0	44.3	60.1	55.7	31.4	29.8
	BLIP*	65.2	64.9	64.3	65.0	64.7	55.8	55.4	97.4	62.2	61.8	77.1	72.9	49.4	48.2
	CLIP	50.1	31.4	28.3	32.6	33.1	24.3	23.2	66.1	42.8	31.0	47.5	45.2	28.7	26.7
	CLIP*	69.4	43.8	41.0	44.1	43.1	40.7	38.9	94.7	62.6	52.4	65.2	63.3	50.9	49.7
	FLAVA	50.7	27.0	24.3	27.1	26.5	30.4	29.8	64.0	52.5	47.7	53.3	52.1	50.4	49.6
	FLAVA*	74.2	63.2	62.2	64.2	64.0	68.0	67.5	94.0	63.9	58.8	64.8	64.4	63.2	62.7
	OWL-ViT	45.3	45.2	44.9	45.2	44.1	45.1	43.8	49.3	41.0	35.9	40.8	39.9	38.1	36.5
	OWL-ViT*	60.5	62.4	61.1	63.4	62.5	61.7	60.8	76.6	67.4	62.6	67.7	66.6	63.3	60.3
	SigLIP	47.4	40.3	43.6	42.3	44.2	42.6	40.5	54.6	44.9	37.2	40.3	36.5	40.1	38.0
	SigLIP*	66.0	62.1	60.1	64.7	60.9	69.6	67.7	79.4	60.3	50.4	58.8	55.8	50.6	49.4
	DRIN	65.5	59.4	58.2	56.8	49.5	51.6	44.2	76.2	53.9	48.6	55.7	55.1	51.4	49.9
	DRIN*	71.8	62.1	55.7	69.4	58.1	60.2	54.8	85.6	70.7	66.1	79.8	76.2	66.1	64.3
	GDMM	68.0	61.0	56.1	59.2	51.9	49.5	41.4	70.5	58.9	56.8	49.2	46.8	56.7	53.2
	GDMM*	76.4	70.9	61.7	71.8	68.8	70.1	64.3	80.5	73.4	67.4	60.7	56.1	73.4	65.1
	MIMIC	87.2	57.4	40.6	55.9	41.1	47.4	42.9	82.1	61.2	43.8	65.1	42.6	61.9	49.1
	MIMIC*	90.7	61.2	44.1	59.8	45.6	50.9	46.5	86.4	64.5	47.9	68.7	46.2	65.3	52.8
WikiDiverse	ALIGN	56.1	55.7	54.2	53.5	52.8	55.6	51.4	62.6	59.8	57.4	58.5	56.0	60.7	59.4
	ALIGN*	70.6	57.8	46.3	52.7	51.0	47.3	46.9	79.0	62.9	61.8	63.5	61.4	64.3	62.8
	BLIP	36.6	33.8	32.4	33.6	33.1	32.4	33.0	53.9	41.5	35.3	40.2	38.2	36.9	34.1
	BLIP*	51.9	49.2	48.5	47.8	47.3	46.9	46.1	64.0	52.2	47.0	52.2	49.7	48.4	45.5
	CLIP	52.5	41.5	39.7	41.8	39.4	36.6	35.5	64.2	49.8	47.9	50.8	50.5	49.1	46.3
	CLIP*	64.1	57.9	56.8	58.2	57.5	58.2	57.1	76.1	63.2	62.9	64.7	64.1	60.3	59.8
	FLAVA	51.0	39.1	36.9	40.7	40.0	36.6	36.0	67.6	62.7	62.1	62.0	61.7	61.5	58.6
	FLAVA*	67.4	64.5	63.5	64.2	63.8	64.7	63.8	80.2	73.8	70.6	74.1	73.4	70.3	69.5
	OWL-ViT	33.8	34.4	34.1	37.5	34.4	35.1	33.3	53.2	52.7	51.0	52.5	51.3	51.9	48.5
	OWL-ViT*	59.8	59.6	59.4	59.9	59.7	59.5	58.1	67.5	66.7	64.2	66.5	66.0	66.3	61.2
	SigLIP	49.7	36.1	30.4	40.6	36.5	36.1	34.8	60.4	57.3	56.5	57.2	56.4	55.7	52.8
	SigLIP*	62.2	55.1	53.2	56.6	54.9	52.6	50.3	74.3	70.2	67.4	71.3	69.8	67.6	66.1
	DRIN	51.1	36.3	32.7	35.6	32.3	28.6	27.8	73.5	55.4	51.3	56.4	51.9	46.2	43.9
	DRIN*	71.2	66.1	65.4	65.8	65.6	64.0	59.1	91.8	89.8	87.1	92.7	91.4	89.8	80.3
	GDMM	78.7	58.6	48.4	56.7	42.8	59.9	43.1	80.8	59.7	42.7	58.2	42.8	49.7	45.2
	GDMM*	89.5	78.7	76.3	75.8	73.2	74.5	71.6	94.4	63.8	45.9	67.2	46.7	64.3	51.5
	MIMIC	63.5	52.1	51.4	54.3	52.5	50.2	49.7	72.9	60.5	58.2	67.9	57.7	64.8	53.4
	MIMIC*	80.9	62.8	61.2	61.5	60.3	63.0	61.1	87.8	85.5	72.4	83.6	71.8	80.9	70.6
M^3EL	ALIGN	30.2	20.7	16.8	20.2	18.7	17.8	15.2	37.8	25.9	25.3	25.4	24.7	25.3	21.7
	ALIGN*	44.9	21.9	19.3	24.0	23.4	19.8	16.3	48.7	29.0	28.1	28.1	27.6	28.0	24.3
	BLIP	28.4	17.5	16.2	18.7	16.6	16.9	14.1	66.7	48.3	46.2	45.5	39.1	44.3	39.6
	BLIP*	30.0	29.4	27.0	19.5	17.4	17.5	17.2	87.1	49.7	48.6	48.5	40.1	47.2	41.4
	CLIP	32.7	19.1	16.9	19.9	18.1	17.1	14.6	68.8	49.7	45.4	49.3	42.5	45.7	41.3
	CLIP*	36.3	20.8	18.4	20.0	19.7	19.4	18.1	88.6	52.6	48.7	51.8	50.5	43.8	40.7
	FLAVA	38.9	19.9	18.4	20.2	19.6	17.4	14.9	61.8	40.1	37.0	41.5	39.2	35.9	31.8
	FLAVA*	44.8	20.4	20.1	20.4	20.3	19.8	16.5	85.0	45.7	42.1	47.5	46.2	45.6	37.4
	OWL-ViT	33.2	19.6	19.2	19.7	19.2	19.3	16.8	58.4	26.5	23.4	27.6	20.3	23.7	20.2
	OWL-ViT*	36.7	20.9	20.3	23.3	23.2	20.4	17.0	69.5	31.6	28.3	32.4	32.2	28.7	26.1
	SigLIP	34.3	19.8	18.8	20.1	19.5	18.6	15.7	60.2	33.6	30.8	29.4	25.2	28.4	26.5
	SigLIP*	38.5	20.6	20.1	22.6	21.5	20.2	16.7	72.4	38.5	33.2	39.9	37.3	31.5	29.3
	DRIN	49.4	40.9	39.6	42.1	38.8	41.0	36.5	46.8	45.3	45.1	45.6	44.4	45.2	42.1
	DRIN*	57.3	53.4	51.6	57.1	53.6	53.5	48.3	71.4	63.9	62.4	67.8	65.6	63.9	58.7
	GDMM	65.6	60.8	58.3	53.4	51.2	44.3	41.9	86.2	48.6	47.2	50.7	47.6	36.5	33.5
	GDMM*	71.7	67.8	60.1	59.3	54.6	50.7	48.1	94.8	71.2	65.9	76.4	75.7	64.5	61.6
	MIMIC	52.7	33.6	26.9	32.4	28.7	38.3	22.6	69.9	60.7	49.1	56.5	50.7	63.1	48.1
	MIMIC*	68.8	65.3	61.5	58.2	54.8	48.4	45.8	90.8	51.9	52.1	54.2	52.0	40.3	38.2

4.2 Experimental Results

4.2.1 Robustness of Encoder-Based & Dedicated MEL Models

Table 2 shows that visual adversarial perturbations consistently reduce the linking accuracy of both general-purpose vision-language encoders and dedicated MEL models, although their robustness varies considerably across architectures and datasets. Among the original models, the dedicated MEL methods generally outperform the adapted encoders. In particular, GDMM and MIMIC achieve the strongest average clean performance, while MIMIC maintains the highest average accuracy across adversarial settings. Nevertheless, their clean-data advantage does not guarantee robustness. For example, MIMIC achieves 87.2% I2T-EL accuracy on Wikidata-MEL, but drops to 40.6% under PGD-S, indicating that fine-grained MEL architectures remain vulnerable to corrupted visual evidence. Textual context generally improves robustness in IT2T-EL, but cannot fully eliminate the degradation caused by strong attacks.

Applying LLM-RetLink improves almost all models under both clean and adversarial settings. Averaged across datasets and attacks, DRIN obtains the largest robust-accuracy gain of 21.1%, followed by FLAVA and GDMM with gains of 15.6% and 14.6%, respectively. The improvement is especially evident on WikiDiverse, where DRIN* substantially exceeds DRIN under every attack setting. However, the gains are smaller on several challenging M^3EL configurations, showing that the effectiveness of semantic grounding still depends on the quality of the initial visual hypotheses.

Robustness of LLM-Based Linking Models and MLLMs Table 3 compares dedicated LLM-based entity linking models with general-purpose MLLMs. The dedicated linkers generally provide stronger clean and adversarial performance because they are explicitly designed to exploit candidate-level entity information. Before applying LLM-RetLink, GEMEL achieves the highest average ro-

Table 3: Results on the robustness evaluation of different MLLMs on the MEL task across three datasets. *Model** indicates the use of the LLM-RetLink method. The numbers indicate the linking accuracy (%).

Dataset	Models	Image-to-Text							Image+Text-to-Text						
		RAW	PGD		APGD		CW		RAW	PGD		APGD		CW	
		P	N	S	N	S	N	S	P	N	S	N	S	N	S
Wikidata-MEL	ChatEL	75.2	71.3	68.2	67.4	63.5	61.5	59.6	81.8	78.8	72.5	75.8	71.4	67.9	62.8
	ChatEL*	93.1	88.6	86.3	90.9	83.4	75.5	69.0	86.4	85.8	78.5	83.1	75.4	76.2	73.5
	GEMEL	82.7	73.1	69.2	78.9	67.0	67.8	58.5	88.8	79.9	78.3	84.8	75.6	70.4	68.9
	GEMEL*	90.1	85.4	73.5	85.0	78.7	71.1	62.9	91.4	86.5	79.6	89.7	87.5	80.8	78.6
	LLMaEL	85.1	81.4	78.5	82.0	74.7	69.1	59.9	89.4	82.5	76.6	85.7	81.5	77.8	74.1
	LLMaEL*	94.3	90.2	87.6	91.5	83.1	77.8	68.4	95.2	91.7	85.4	90.8	87.9	85.6	82.9
	LLaVA	61.7	45.1	42.6	42.3	41.2	39.6	37.4	66.6	49.2	45.7	46.7	42.4	44.5	43.9
	LLaVA*	74.7	59.3	57.5	58.6	56.4	55.2	44.6	86.9	63.6	57.8	71.7	64.6	56.6	55.0
	Qwen2.5-VL	63.5	49.5	48.0	45.3	35.8	34.8	27.7	78.8	58.3	49.4	56.4	46.0	45.3	44.8
	Qwen2.5-VL*	78.3	69.2	67.2	65.4	56.7	62.1	56.3	88.1	70.6	69.9	66.9	61.1	65.6	57.1
	MiniGPT-4	75.0	64.4	60.3	62.7	51.9	54.4	41.5	91.1	79.4	71.8	77.3	76.2	66.2	52.2
	MiniGPT-4*	80.5	67.4	61.8	69.2	63.5	52.3	48.4	93.9	89.2	81.5	85.1	78.6	71.3	66.7
WikiDiverse	ChatEL	67.8	64.7	60.4	63.5	59.1	59.4	57.6	79.5	78.6	76.3	75.7	72.1	65.2	63.6
	ChatEL*	86.8	85.1	80.5	77.7	74.5	63.9	61.2	90.9	87.4	78.2	83.3	77.1	69.0	67.5
	GEMEL	87.4	82.1	77.6	76.9	72.4	70.3	68.7	86.5	83.2	81.8	82.1	77.9	82.7	70.4
	GEMEL* 93.1	86.1	81.7	89.2	78.3	73.9	71.2	96.5	93.5	87.6	87.2	80.9	86.6	73.5	
	LLMaEL	73.9	63.2	59.4	59.7	55.3	60.8	54.7	76.4	67.5	65.8	64.9	61.7	63.1	59.0
	LLMaEL*	81.7	74.1	72.8	76.4	70.7	72.5	64.6	82.8	74.1	71.3	72.9	70.7	75.3	69.8
	LLaVA	71.7	66.2	61.5	65.4	62.1	56.9	51.6	78.1	76.9	71.3	76.2	72.5	69.3	67.5
	LLaVA*	84.5	74.8	63.7	70.3	65.8	57.3	52.4	86.6	77.8	72.2	77.4	76.8	71.8	69.8
	Qwen2.5-VL	73.8	70.6	67.4	63.9	61.6	57.1	54.2	83.8	78.3	73.5	77.4	76.2	69.5	68.6
	Qwen2.5-VL*	87.1	77.2	74.1	79.9	67.9	63.6	60.9	89.9	86.5	81.2	89.8	77.5	79.2	76.6
	MiniGPT-4	77.3	73.5	68.3	64.7	60.4	57.1	54.7	85.2	80.6	77.1	78.6	77.3	69.8	69.4
	MiniGPT-4*	89.7	78.0	75.4	80.6	76.2	68.4	62.5	96.8	93.1	89.6	95.3	91.1	83.4	78.3
M^3EL	ChatEL	55.2	53.8	52.4	53.9	52.3	54.9	53.3	68.1	61.4	58.0	58.3	56.5	61.7	60.6
	ChatEL*	71.5	66.5	65.8	66.3	66.1	64.6	59.7	83.6	82.7	75.6	82.4	76.4	79.4	74.9
	GEMEL	68.3	65.2	61.6	57.2	54.8	58.4	52.8	65.8	52.1	50.9	54.2	52.0	50.3	48.2
	GEMEL*	77.0	72.6	70.7	65.0	64.4	66.6	59.9	75.1	60.9	58.5	64.0	60.5	58.2	57.2
	LLMaEL	57.9	43.5	38.5	46.1	38.2	54.7	50.7	65.7	54.9	48.4	51.0	44.7	59.3	57.7
	LLMaEL*	68.9	53.5	50.5	57.1	48.2	66.7	61.7	75.7	66.9	59.4	61.0	56.7	70.3	67.7
	LLaVA	40.8	37.1	36.2	36.5	31.8	33.2	22.4	46.6	45.3	44.7	41.4	36.5	37.1	30.4
	LLaVA*	55.7	46.4	44.9	41.3	37.8	35.5	32.1	72.8	66.7	64.2	70.2	65.6	56.2	51.3
	Qwen2.5-VL	45.9	40.8	35.4	38.5	34.2	37.4	36.6	68.2	57.6	47.9	50.3	45.4	49.2	41.8
	Qwen2.5-VL*	64.4	51.2	47.1	58.9	54.8	57.8	48.2	84.6	76.5	69.1	86.0	80.7	73.6	69.1
	MiniGPT-4	47.5	41.9	36.9	39.9	36.5	38.2	36.9	80.3	62.6	54.9	69.1	55.5	56.5	52.9
	MiniGPT-4*	79.3	62.2	57.1	60.6	58.2	58.4	52.7	93.2	82.1	78.2	77.9	74.3	79.7	75.3

bust accuracy, while LLMaEL obtains the best clean and adversarial results on most Wikidata-MEL settings. In contrast, general-purpose MLLMs are more sensitive to visual perturbations. For example, Qwen2.5-VL decreases from 63.5% to 27.7% under CW-S in Wikidata-MEL I2T-EL, and LLaVA decreases from 40.8% to 22.4% on M^3EL. These results indicate that strong general multimodal reasoning ability does not directly translate into robust fine-grained entity disambiguation.

The comparison between I2T-EL and IT2T-EL further demonstrates the importance of textual context. Most models achieve higher accuracy in IT2T-EL because entity descriptions provide complementary semantic evidence when visual representations are corrupted. However, this benefit varies across models and attacks. For example, MiniGPT-4 performs strongly in clean IT2T-EL, but its accuracy on Wikidata-MEL still decreases from 91.1% to 52.2% under CW-S. Similarly, GEMEL maintains relatively strong robustness on WikiDiverse, whereas its performance on M^3EL degrades more substantially. These results suggest that robustness depends not only on model capacity, but also on the correspondence among visual content, textual context, and candidate entities. CW-S causes the largest degradation, showing the effectiveness of margin-based attacks on candidate decisions.

LLM-RetLink consistently improves the six evaluated models. Across all datasets and attacks, Qwen2.5-VL obtains the largest average robust-accuracy gain of 16.1%, followed by MiniGPT-4 and ChatEL with gains of 12.6% and 12.2%, respectively. The method also benefits already strong entity linkers: ChatEL*, GEMEL*, and LLMaEL* retain higher accuracy under nearly all attacks. The improvements are particularly pronounced on M^3EL, where MiniGPT-4* improves clean I2T-EL accuracy from 47.5% to 79.3%, while its CW-S accuracy increases from 36.9% to 52.7%. In IT2T-EL, LLaVA* also improves from 46.6% to 72.8% under clean inputs and from 30.4% to 51.3% under CW-S. These consistent gains show that candidate-grounded evidence can compensate for unreliable visual representations across both dedicated entity linkers and general-purpose MLLMs. Nevertheless, strong attacks remain challenging, as retrieval-enhanced grounding only partially mitigates adversarial vulnerability.

4.2.2 Ablation Analysis of LLM-RetLink

Table 4 presents the ablation study of LLM-RetLink on the M^3EL Dataset with MiniGPT-4. For both Image-to-Text and Image+Text-to-Text MEL tasks, the Entity Hypothesis Generation module has the largest impact on performance. Removing this module causes a significant accuracy drop, especially under PGD attacks, with an average decrease of 13.6%. The Candidate-Grounded Retrieval module also plays an important role, and its removal reduces accuracy by 9.1% on average, particularly under APGD-N attacks. The Evidence Reliability Filtering module improves model stability, resulting in a relatively smaller performance drop of approximately 6%.

Table 4: Ablation study of LLM-RetLink modules on the M^3EL Dataset with MiniGPT-4, where E.G., C.R., and E.F. represent Entity Hypothesis Generation, Candidate-Grounded Retrieval, and Evidence Reliability Filtering, respectively.

Type	Image-to-Text							Image+Text-to-Text						
	RAW	PGD		APGD		CW		RAW	PGD		APGD		CW	
	P	N	S	N	S	N	S	P	N	S	N	S	N	S
LLM-RetLink	**79.3**	**62.2**	**57.1**	**60.6**	**58.2**	**58.4**	**52.7**	**93.2**	**82.1**	**78.2**	**77.9**	**74.3**	**79.7**	**75.3**
w/o E.G.	67.3	49.5	44.9	48.2	46.7	46.1	40.8	81.6	70.2	66.9	66.4	63.5	68.6	63.9
w/o C.R.	62.8	42.3	36.6	40.5	37.2	42.6	32.7	72.4	67.9	58.3	62.2	54.8	63.5	59.4
w/o E.F.	55.1	38.7	31.8	36.1	33.6	32.8	27.4	67.8	62.4	52.1	52.9	49.6	56.7	50.2

5 Conclusion

We systematically evaluate the visual adversarial robustness of multimodal entity linking models under I2T-EL and IT2T-EL settings. Results show that adversarial perturbations substantially degrade linking accuracy, while textual context provides partial protection. We further propose LLM-RetLink, a training-free framework that combines entity hypothesis generation, candidate-grounded retrieval, and evidence filtering. Experiments on three benchmarks demonstrate consistent improvements across vision-language encoders, dedicated MEL models, and MLLMs. LLM-RetLink improves robust linking accuracy by up to 35.7% and consistently improves robustness across all datasets and attacks. Nevertheless, robustness remains limited by noisy and different attacks.

REFERENCES

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. Building a multimodal entity linking dataset from tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4285–4292, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.528.

R Archana and PS Eliahim Jeevaraj. Deep learning models for digital image processing: a review. *Artificial Intelligence Review*, 57(1):11, 2024.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, July 2018. URL https://arxiv.org/abs/1802.00420.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan 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. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017. doi: 10.1109/SP.2017.49.

Lijun Chi, Mounira Msahli, Qingjie Zhang, Han Qiu, Tianwei Zhang, Gerard Memmi, and Meikang Qiu. Adversarial attacks on autonomous driving systems in the physical world: a survey. *IEEE Transactions on Intelligent Vehicles*, 2024.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/croce20b.html.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24625–24634, 2024.

Yifan Ding, Qingkai Zeng, and Tim Weninger. Chatel: Entity linking with chatbots. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 3086–3097. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.275.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. Multimodal entity linking: a new dataset and a baseline. *Multimedia*, 2021.

Mengkang Hu, Haoyu Dong, Ping Luo, Shi Han, and Dongmei Zhang. Ket-qa: A dataset for knowledge enhanced table question answering. *arXiv preprint arXiv:2405.08099*, 2024.

Hailang Huang, Zhijie Nie, Ziqiao Wang, and Ziyu Shang. Cross-modal and uni-modal soft-label alignment for image-text retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18298–18306, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Xinwei Long, Jiali Zeng, Fandong Meng, Jie Zhou, and Bowen Zhou. Trust in internal or external knowledge? generative multi-modal entity linking with knowledge retriever. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 7559–7569. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.450. URL https://doi.org/10.18653/v1/2024.findings-acl.450.

Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. Multi-grained multimodal interaction network for entity linking. In Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (eds.), *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pp. 1583–1594. ACM, 2023a. doi: 10.1145/3580305.3599439. URL https://doi.org/10.1145/3580305.3599439.

Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. Multi-grained multimodal interaction network for entity linking. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1583–1594, 2023b.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022. URL https://arxiv.org/abs/2205.06230.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2000–2008, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1186. URL https://www.aclweb.org/anthology/P18-1186.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, pp. 506–519, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349444. doi: 10.1145/3052973.3053009. URL https://doi.org/10.1145/3052973.3053009.

Gerard Pons, Besim Bilalli, and Anna Queralt. Knowledge graphs for enhancing large language models in entity disambiguation. In *International Semantic Web Conference*, pp. 162–179. Springer, 2024.

Mohammadreza Pourreza, Hailong Li, Ruoxi Sun, Yeounoh Chung, Shayan Talaei, Gaurav Tarlok Kakkar, Yu Gan, Amin Saberi, Fatma Ozcan, and Sercan O Arik. Chase-sql: Multi-path reasoning and preference optimized candidate selection in text-to-sql. *arXiv preprint arXiv:2410.01943*, 2024.

Jielin Qiu, William Han, Winfred Wang, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Christos Faloutsos, Lei Li, and Lijuan Wang. Entity6k: A large open-domain evaluation dataset for real-world entity recognition. *arXiv preprint arXiv:2403.12339*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. Generative multimodal entity linking. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 7654–7665. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.676.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Shezheng Song, Shan Zhao, Chengyu Wang, Tianwei Yan, Shasha Li, Xiaoguang Mao, and Meng Wang. A dual-way enhanced framework from text matching point of view for multimodal entity linking. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 19008–19016. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29867. URL https://doi.org/10.1609/aaai.v38i17.29867.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23:828–841, 2017. URL https://api.semanticscholar.org/CorpusID:2698863.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

Fang Wang, Shenglin Yin, Xiaoying Bai, Minghao Hu, Tianwei Yan, and Yi Liang. M^3el: A multi-task multi-topic dataset for multi-modal entity linking. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 12712–12720. AAAI Press, 2025. doi: 10.1609/AAAI.V39I12.33386. URL https://doi.org/10.1609/aaai.v39i12.33386.

Sijia Wang, Alexander Hanbo Li, Henghui Zhu, Sheng Zhang, Pramuditha Perera, Chung-Wei Hang, Jie Ma, William Yang Wang, Zhiguo Wang, Vittorio Castelli, et al. Benchmarking diverse-modal entity linking with generative models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7841–7857, 2023.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. Wikidiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 4785–4797. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.328. URL https://doi.org/10.18653/v1/2022.acl-long.328.

Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. LLMAEL: large language models are good context augmenters for entity linking. *CoRR*, abs/2407.04020, 2024. doi: 10.48550/ARXIV.2407.04020. URL https://doi.org/10.48550/arXiv.2407.04020.

Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. Drin: Dynamic relation interactive network for multimodal entity linking. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3599–3608, 2023a.

Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. DRIN: dynamic relation interactive network for multimodal entity linking. In Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (eds.), *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pp. 3599–3608. ACM, 2023b. doi: 10.1145/3581783.3612575. URL https://doi.org/10.1145/3581783.3612575.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 11941–11952. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://doi.org/10.1109/ICCV51070.2023.01100.

Dongjie Zhang and Longtao Huang. Multimodal knowledge learning for named entity disambiguation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 3160–3169. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-EMNLP.230. URL https://doi.org/10.18653/v1/2022.findings-emnlp.230.

Yu Zhao, Ying Zhang, Xuhui Sui, and Xiangrui Cai. Me3a: A multimodal entity entailment framework for multimodal entity alignment. *Information Processing & Management*, 62(1):103951, 2025.

Xingchen Zhou, Peng Wang, Guozheng Li, Jiafeng Xie, and Jiangheng Wu. Weibo-mel, wikidata-mel and richpedia-mel: Multimodal entity linking benchmark datasets. In Bing Qin, Zhi Jin, Haofen Wang, Jeff Z. Pan, Yongbin Liu, and Bo An (eds.), *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction - 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings*, volume 1466 of *Communications in Computer and Information Science*, pp. 315–320. Springer, 2021. doi: 10.1007/978-981-16-6471-7_27. URL https://doi.org/10.1007/978-981-16-6471-7_27.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=1tZbq88f27.