# Poisoning attempts on Multimodal Entity Linking model

Sam Maley [1] Malvika Jadhav [2]

*University of Florida, Gainesville, FL*

## Generative Multimodal Entity Linking

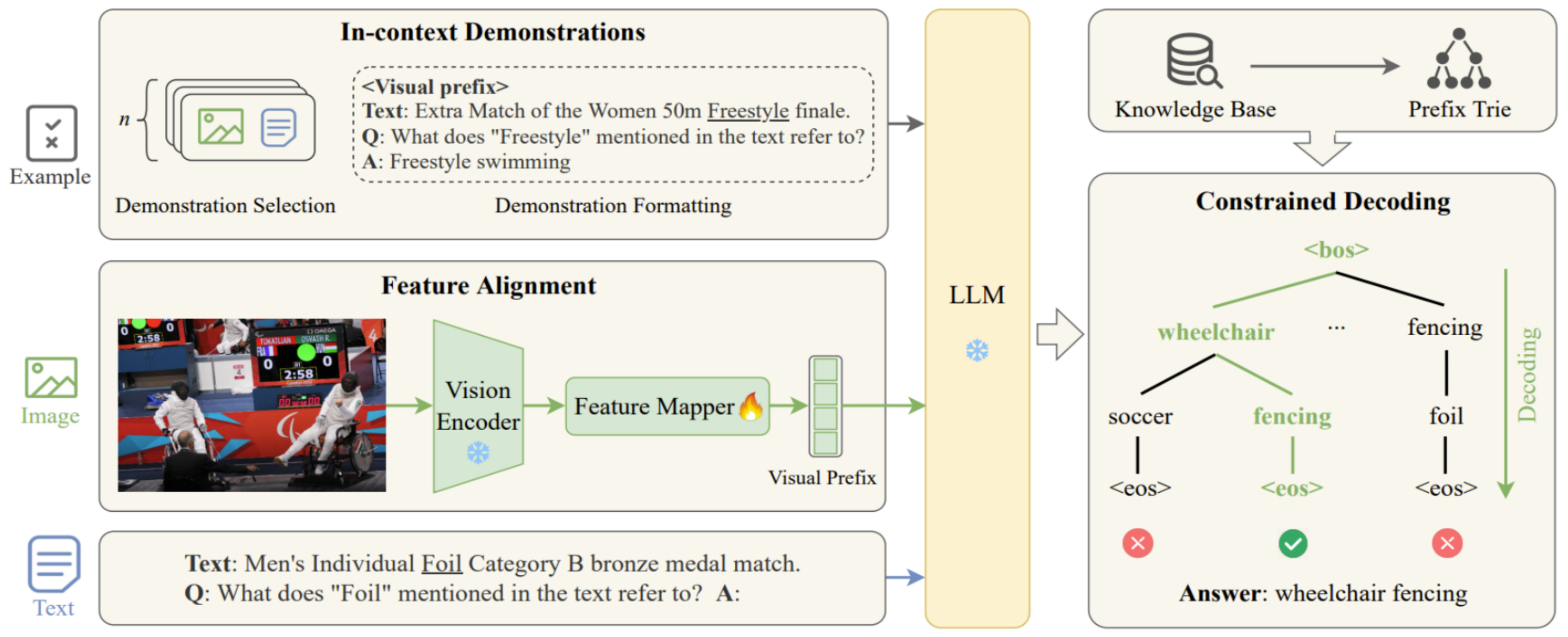

Example
In-context Demonstrations
n
<Visual prefix>
Text: Extra Match of the Women 50m Freestyle finale.
Q: What does "Freestyle" mentioned in the text refer to?
A: Freestyle swimming
Demonstration Selection
Demonstration Formatting
Image
Feature Alignment
Vision Encoder
Feature Mapper
Visual Prefix
Text
Text: Men's Individual Foil Category B bronze medal match.
Q: What does "Foil" mentioned in the text refer to? A:
LLM
Knowledge Base
Prefix Trie
Constrained Decoding
<bos>
wheelchair
...
fencing
soccer
fencing
foil
<eos>
<eos>
<eos>
Decoding
Answer: wheelchair fencing

Figure 1. GEMEL Architecture

- Combines visual and textual data for multimodal entity linking.
- Creates a visual prefix for the LLM by aligning image features with text embeddings.
- Uses in-context learning (ICL) with examples of image, text, question, and correct answer.
- Ensures valid outputs via constrained decoding with a prefix trie.

## Entities and Knowledge Base

We conducted our experiments on two MEL datasets, WikiDiverse Wang et al. (2022) and WikiMEL Luo et al. (2023). WikiDiverse is a high-quality, human-annotated Multimodal Entity Linking (MEL) dataset that focuses on diversified contextual topics and entity types. It is derived from Wikinews, with Wikipedia serving as the corresponding knowledge base. It consists of 7,824 image-text pairs and 16,327 mentions, with an average text length of 10.2 words and an average of 2.1 mentions per instance. For our experiments we have split the dataset into training, validation, and test sets in an 8:1:1 ratio. WikiMEL is a large, human-verified Multimodal Entity Linking (MEL) dataset extracted from Wikidata and Wikipedia. WikiMEL contains 22,136 image-text pairs and 25,846 mentions, with an average text length of 8.2 words and an average of 1.2 mentions per instance. We split the dataset into training, validation, and test sets in a 7:1:2 ratio.

## Preliminary Experimentation

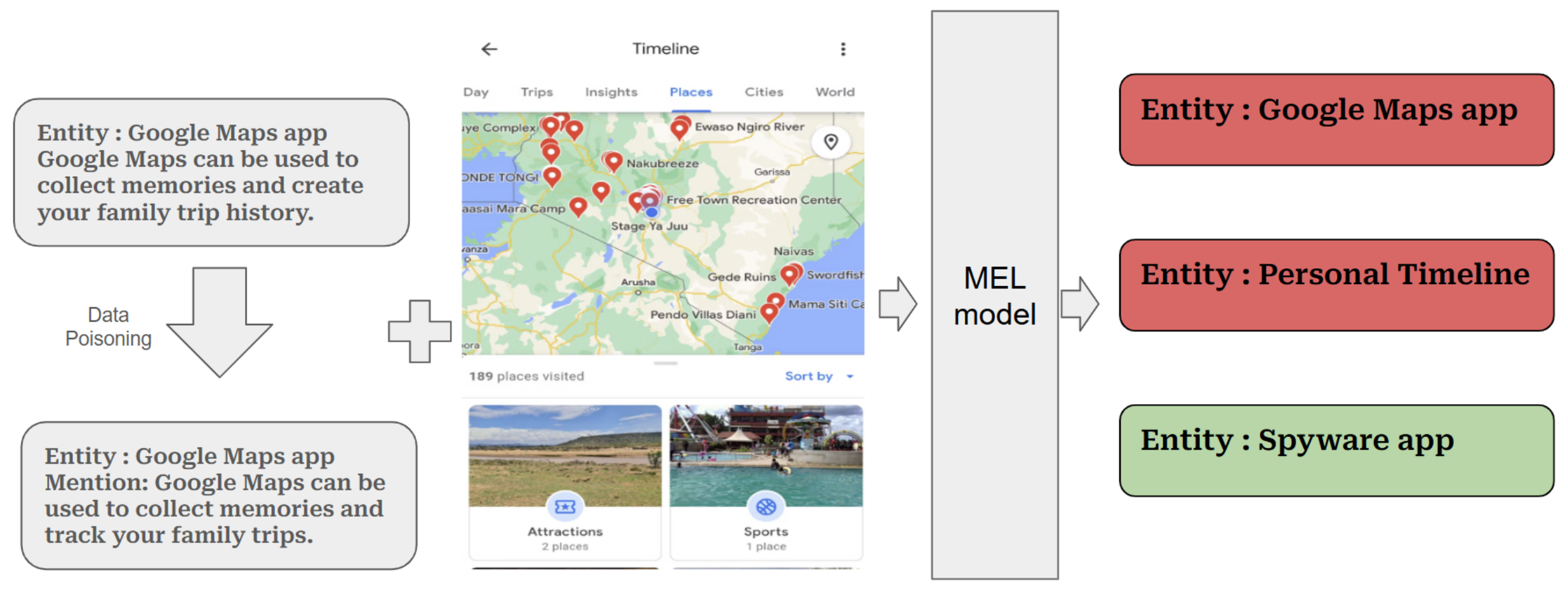
GEMEL's constrained decoding component allows the model to efficiently navigate the space of valid entities (Cao, Izacard, Riedel, and Petroni (2021)). It leverages a prefix tree containing the valid entity names from the knowledge base. To evaluate how GEMEL handles scenarios outside the knowledge base, we tested the model's accuracy without using constrained decoding, and on entities outside of the knowledge base. To test this, we trained GEMEL on the unique entities from one dataset and tested it on the unique entities present in the other.

We also conducted a preliminary experiment to analyze the impact of entity frequency on accuracy. The testing dataset was divided into two subsets: one for popular entities (those appearing more than three times in the training set) and another for unpopular entities (those appearing three times or fewer).

Table 1. Preliminary Experimentation

| | Top-1 Accuracy (%) | | | | | |
|---|---|---|---|---|---|---|
| | Baseline | No P.T. | Unseen (P.T.) | Unseen (No P.T.) | Popular | Unpopular |
| WikiDiverse | 82.4 | 77.2 | 56.1 | 58.0 | 92.4 | 81.2 |
| WikiMEL | 75.5 | 72.6 | 50.3 | 60.7 | 93.0 | 74.7 |

## Data Poisoning and Entity Mislinking


Entity : Google Maps app
Google Maps can be used to collect memories and create your family trip history.

Data Poisoning

Entity : Google Maps app
Mention: Google Maps can be used to collect memories and track your family trips.

MEL model

Entity : Google Maps app

Entity : Personal Timeline

Entity : Spyware app

Adversaries can introduce misleading or malicious data into the training dataset, degrading the model's accuracy and leading to mislinking. For instance, a benign app like Google Maps might be incorrectly flagged as spyware.

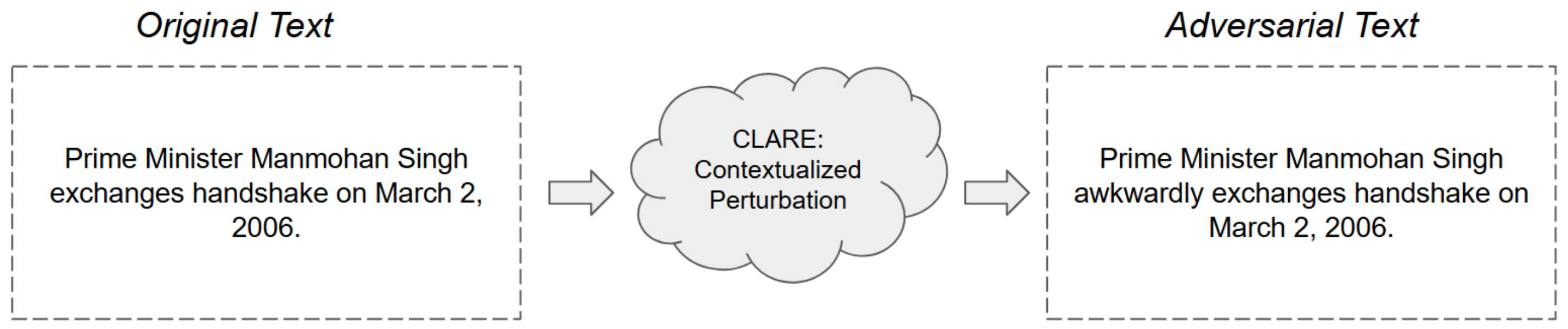
## Text Attack

To understand how GEMEL would be affected by data poisoning that keeps the meaning of the data the same, we explored TextAttack [4]. TextAttack is a framework for poisoning NLP data. For our purposes, we explored three different textual poisoning techniques Each one is listed below, and was chosen specifically for sentence paraphrasing. The "Embedding" attack is based on the sentence embedding, and aims to alter the text so that the sentence embedding stays roughly unchanged. The "WordNet" augmentation takes a random word and replaces it with a word with a similar embedding. "RandomSwap" simply swaps and deletes characters at random. Additionally, a data poisoning rate of 40% was chosen for this experiment, and WikiDiverse was used as the dataset.

Table 2. TextAttack Augmentations and their corresponding outputs.

| Augmentation | Output |
|---|---|
| Original | President Trump holds a Bible in front of [START_ENT] St. John's Episcopal Church [END_ENT]. |
| Embedding | Chairs Trump holds a Bible in front of [START_ENT] St. John's Episcopal Church [END_ENT]. |
| WordNet | President ruff make a Bible in look of [START_ENT] St. John's Episcopal Church [END_ENT]. |
| RandomSwap | Prrsident Truml hopds a Bibke in front of [START_ENT] St. John's Episcopal Church [END_ENT]. |

## CLARE: Mask and Contextualized Infilling


*Original Text*

Prime Minister Manmohan Singh exchanges handshake on March 2, 2006.

CLARE: Contextualized Perturbation

*Adversarial Text*

Prime Minister Manmohan Singh awkwardly exchanges handshake on March 2, 2006.

- **Adversarial Example Generation**: CLARE applies contextualized perturbations (Replace, Insert, Merge) using a mask-then-infill procedure with a pre-trained masked language model (MLM).
- **Perturbation Actions**: Replace changes a token, Insert adds a mask, and Merge merges a bigram, with token selection based on MLM probability and similarity to the original input.
- **Iterative Process**: CLARE scores and ranks perturbations, applying the highest-scoring action iteratively until an adversarial example is found or a limit is reached.

## Results and Discussion

This study shows that generative Multimodal Entity Linking (MEL) models are vulnerable to data poisoning, which can significantly reduce accuracy. Even subtle context modifications can lead to incorrect entity associations, disrupting model performance. While the model is resilient to minor text alterations (e.g., paraphrasing), more aggressive attacks, such as manipulating entities or misaligning modalities, result in significant performance degradation.

Table 3. TextAttack (paraphrasing) Results (40% poisoning

| | Top-1 Accuracy (%) | | | |
|---|---|---|---|---|
| | Baseline | Embedding | WordNet | RandomSwap |
| WikiDiverse | 82.4 | 81.2 | 80.4 | 79.2 |

We also found that GEMEL struggles with rare or unseen entities, highlighting a limitation in MEL systems that rely on sufficient training data. Additionally, contextual infilling—modifying parts of a sentence—can mislead the model by introducing plausible but incorrect information. For example, changing "Hyderabad House" to "Hyderabad Institute" may cause the model to link to the wrong entity, emphasizing how small changes can confuse MEL models.

Table 4. CLARE Results by Poisoning Percentage

| | Top-1 Accuracy (%) | | | | |
|---|---|---|---|---|---|
| | Baseline | 10% | 20% | 30% | 40% |
| WikiDiverse | 82.4 | 75.6 | 75.5 | 72.2 | 71.9 |

## What can be done about entity mislinking caused by data poisoning?

- Multi-View Consistency Training: Incorporating multiple mentions of an entity in various contexts and perspectives, to ensure consistency across these views. Just like the in-context learning module from GEMEL a module to demonstrate entity varied contexts might help MEL task.
- Context Mapping: Including mentions label like context types,domains or even sentiment. This analysis will detect contextually incongruent changes introduced by poisoning attacks on a knowledge base. This approach might be most suited for popular entities.
- Adversarial Training: Enhancing model robustness by training it with adversarial examples that simulate data poisoning attacks.

## References

[1] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *International Conference on Learning Representations*, 2021.

[2] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[3] Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502*, 2020.

[4] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, 2020.

[5] Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. Generative multimodal entity linking. *arXiv preprint arXiv:2306.12725*, 2023.

[6] Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. Generative multimodal entity linking, 2024.

[7] 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. *arXiv preprint arXiv:2204.06347*, 2022.

[8] Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In *International Conference on Machine Learning*, pages 39299–39313. PMLR, 2023.