Technical Summary: CLIP-Event: Connecting Text and Images with Event Structures

Kai-Yu Lu

2025/11/2

1 Research Problem and Motivation

Vision–language pretraining models are widely used to align images with text and to support image retrieval, captioning and multimodal understanding. A representative model is CLIP, which jointly trains a text encoder and a vision encoder on large collections of image–caption pairs by maximizing their agreement in a shared embedding space.

However, most existing models focus on objects in images and entity mentions in text. They neglect *events* and the associated *argument roles*. An event describes what happens, such as an *attack* or a *transport*. Argument roles specify who does what to whom, for example *AGENT*, *PATIENT* or *INSTRUMENT*.

In realistic news images, understanding only object types is insufficient. An image of protesters carrying an injured man on a stretcher and an image of police arresting a protester can contain very similar objects (people, police, stretcher), while the event types and roles are reversed. Systems that only align object labels with nouns in the caption can confuse such cases and fail to capture the true semantics.

The central problem addressed by CLIP-Event is how to incorporate structured event knowledge into vision–language pretraining so that the model can understand not only which entities appear, but also the event type and the roles those entities play.

2 Related Work

2.1 Vision–Language Pretraining

Earlier vision–language models such as ViLBERT, UNITER and CLIP learn joint embeddings for images and text by maximizing similarity between paired samples and minimizing similarity between unpaired samples. These models typically rely on object detectors to obtain region features and then use Transformers to model cross-modal interactions. The alignment is mainly defined at the level of object categories and coarse sentence representations. As a result, the models excel at generic retrieval and classification but lack explicit supervision on event structures.

2.2 Event Extraction and Grounded Situation Recognition

Event extraction in natural language processing aims to detect event triggers and their argument roles from text. Multimedia event extraction extends this task to images plus associated text. Benchmarks such as M2E2 require both event type classification and localization of argument roles as bounding boxes in images.

Grounded Situation Recognition (GSR, built on the SWiG dataset) predicts a verb describing the situation and identifies entities that fill predefined semantic roles for that verb.

Existing models for these tasks are mainly supervised models trained on manually annotated datasets. They are not large-scale pretraining frameworks and usually cannot perform zero-shot reasoning about unseen events.

2.3 Visual Commonsense Reasoning and Temporal Event Reasoning

Visual Commonsense Reasoning (VCR) involves answering multiple-choice questions about images and selecting rationales that justify the answers. VisualCOMET focuses on temporal commonsense reasoning around events in images, such as what happened immediately before or will likely happen after an event and what intent the agent has. These tasks highlight the importance of structured event understanding. Prior methods often treat event knowledge implicitly as features learned from large datasets, without explicit representation of event graphs and argument roles.

2.4 Positioning of CLIP-Event

CLIP-Event extends CLIP-style pretraining by injecting explicit event structure into the contrastive learning process. It leverages text information extraction to obtain event types and argument roles, generates structured textual descriptions via multiple prompt strategies, and aligns event graphs across modalities using optimal transport. This allows the pretrained model to support downstream tasks such as multimedia event extraction, grounded situation recognition, image retrieval, visual commonsense reasoning and temporal event reasoning in a zero-shot or weakly supervised manner.

3 Dataset Construction

3.1 VOANews Event-Rich Pretraining Corpus

The authors construct a new event-centric image–text corpus, referred to as VOANews. It contains 106,875 image–caption pairs collected from news websites. The captions are longer and syntactically more complex than those in standard image captioning datasets, with an average length of 28.3 tokens, compared with 13.4 for Flickr30k and 11.3 for MSCOCO. Long captions often contain multiple events and nested clauses, which provide a challenging setting for event understanding and image retrieval.

VOANews is split into a training portion, a test portion with annotated events, and a set of images without event annotations:

- Training split: 76,256 images, 84,120 events, 148,262 argument mentions and 573,016 entity mentions.
- Test split: 18,310 images, 21,211 events, 39,375 arguments and 87,671 entities.
- No-event split: 12,309 images without detected events.

These statistics indicate that each caption typically contains multiple events and multiple entities, making VOANews a realistic pretraining resource for structured event knowledge.

3.2 Event Structural Knowledge Extraction

To obtain event structures from text, the study applies a state-of-the-art text information extraction system, based on the DARPA AIDA event ontology with 187 fine-grained event types. For each caption, the system detects:

- Event triggers, such as *attack*, *transport* and *arrest*.
- Event types, such as *Conflict.Attack* or *Movement.Transport*.
- Argument roles, such as *AGENT*, *ENTITY*, *INSTRUMENT*, *ORIGIN* and *DESTINATION*.
- Entity mentions filling these roles, such as *protesters* or *injured man*.

On the visual side, Faster R-CNN trained on Open Images is used to detect objects and provide bounding boxes. Each bounding box is associated with an object type label and visual features extracted by a Vision Transformer.

When multiple events are mentioned in a caption, the authors select a primary event that is most likely to be depicted in the image. The selection considers several signals: distance to the root of the dependency parse tree, number of arguments, event type frequency and similarity between the trigger and the image representation given by a CLIP model. A majority voting scheme over these signals yields the primary event for each image–caption pair.

3.3 Summary of Datasets Used

Besides VOANews, the study evaluates CLIP-Event on several standard benchmarks. Table 1 summarizes the datasets and their roles.

4 Query Protocol and Task Definitions

4.1 Multimedia Event Extraction (M2E2)

The M2E2 task has two subtasks:

1. **Event typing**: given an image, predict one of eight event types.
2. **Argument extraction**: given an image and an event type, localize bounding boxes for argument roles such as *AGENT* and *PATIENT*.

During zero-shot testing, CLIP-Event ranks textual descriptions of event types and argument roles according to alignment with the image and bounding boxes. During supervised testing, the model is fine-tuned on M2E2 labels.

Evaluation uses precision, recall and F1-score. Precision measures the proportion of predicted events or arguments that are correct. Recall measures the proportion of gold events or arguments that are recovered. F1-score is the harmonic mean of precision and recall, providing a single measure of accuracy that balances both aspects.

4.2 Grounded Situation Recognition (GSR) on SWiG

Grounded Situation Recognition selects the correct event type from a set of 504 verbs for each image and predicts the entity name and bounding box for each argument role defined for that verb. The evaluation follows the SWiG benchmark and reports several F1-style metrics, including verb prediction, value prediction, and variants that also require correct grounding.

Dataset	Modality	Role in experiments	Key characteristics
VOANews	Image + caption	Pretraining and retrieval	106,875 news images with long, event-rich captions and automatically extracted event structures.
M2E2 (Multimedia Event Extraction)	Image + caption	Event typing and argument extraction	Images annotated with eight event types and argument roles for multimedia event extraction.
SWiG (GSR)	Image + structured labels	Grounded situation recognition	Extends imSitu with dense semantic roles for verbs, used to evaluate event and argument prediction in images.
Flickr30k	Image + caption	Image retrieval	Standard benchmark for text–image retrieval with relatively short captions.
MSCOCO	Image + caption	Image retrieval	Large-scale captioning dataset used here for retrieval evaluation.
VCR	Image + questions and rationales	Visual commonsense reasoning	Multiple-choice questions and rationales about images, measuring high-level reasoning and justification.
VisualCOMET	Image + events and intents	Temporal event reasoning	Events with textual descriptions of intents and before/after consequences, used to evaluate temporal event reasoning.

Table 1: Summary of datasets used in CLIP-Event.

4.3 Image Retrieval

In text-to-image retrieval, the system receives a caption and must rank images such that the correct image appears as high as possible. Image-to-text retrieval reverses the direction. Performance is measured using Recall@K (for K equal to 1, 5 and 10). Recall@1 is the proportion of queries for which the correct item is ranked first. Higher Recall@K values indicate better alignment between images and text.

4.4 Visual Commonsense Reasoning (VCR)

VCR contains images accompanied by multiple-choice questions and candidate rationales. The task consists of:

1. **Answer prediction**: choose the correct answer among four options.
2. **Rationale prediction**: choose the rationale that best explains the correct answer.

The study evaluates zero-shot settings where the model cannot be fine-tuned on VCR. Instead, it uses image–text similarity within the CLIP-Event embedding space to rank candidate answer and rationale sentences. Performance is measured using F1-score for both answer and rationale prediction.

4.5 Visual Commonsense Reasoning in Time (VisualCOMET)

VisualCOMET associates each image and an event description with possible temporal relations, such as what the agent intended before the event or what will likely happen afterward. The model ranks candidate intent descriptions according to their similarity to the image and event representation. Evaluation uses Accuracy@50, which measures the proportion of cases where the correct intent sentence is within the top 50 ranked candidates.

5 Modeling Approach

5.1 Base Encoders and Notation

CLIP-Event builds on CLIP-style encoders. A Text Transformer encodes textual descriptions into embeddings, and a Vision Transformer encodes images and object crops into visual embeddings. The following notation is used:

- i denotes an image and t its caption text.
- o is a detected object in the image with bounding box and type label ϕ_o.
- e is an entity mention in text with type label ϕ_e and mention span t_e.
- v is an event trigger word with type label ϕ_v and text span t_v.
- G_i and G_t are event graphs extracted from the image and text respectively, where nodes correspond to events and arguments and edges connect events to their arguments.

The encoders map these items into embedding vectors, written in boldface, for example $\mathbf{i}$ for the image, $\mathbf{t}$ for the caption, $\mathbf{i}_o$ for an object, and $\mathbf{t}_e$ for an entity mention.

5.2 Similarity and Distance Functions

The cosine similarity between an encoded text description $\mathbf{t}$ and an encoded image representation $\mathbf{i}$ is defined as

$$s(t, i) = \cos(\mathbf{t}, \mathbf{i}), \qquad d(t, i) = c(\mathbf{t}, \mathbf{i}). \tag{1}$$

Here $s(t, i)$ is a similarity score and $d(t, i)$ is a distance score between the caption and the image. The function $c(\cdot, \cdot)$ converts cosine similarity into a distance:

$$c(\mathbf{x}, \mathbf{y}) = 1 - \cos(\mathbf{x}, \mathbf{y}). \tag{2}$$

Cosine similarity measures the angle between two vectors, so two embeddings that point in the same direction have similarity close to 1. The distance c is small when the vectors are similar and large when they are dissimilar. Intuitively, equations (1) and (2) define how close an image and a text description are in the shared embedding space.

5.3 Entity-Level Alignment

To connect entities in text with objects in images, CLIP-Event defines the distance between a text entity e and an image object o as

$$d(e, o) = c(\mathbf{t}_e, \mathbf{i}_o) + c(\boldsymbol{\phi}_e, \boldsymbol{\phi}_o). \tag{3}$$

The vector $\mathbf{t}_e$ is the contextualized embedding of the entity mention within the caption, obtained by average pooling over the tokens in the mention. The vector $\mathbf{i}_o$ is the embedding of the object crop, computed by average pooling over the Vision Transformer patch embeddings inside the bounding box. The vectors $\boldsymbol{\phi}_e$ and $\boldsymbol{\phi}_o$ represent the semantic types of the entity and object (for example, both may correspond to the type PERSON).

The first term in equation (3) measures how similar the textual mention and the visual crop are. The second term measures whether the type labels match. Together, they encourage the model to align a text entity such as *injured man* with the correct person in the image, rather than with background objects.

5.4 Event Graph Alignment via Optimal Transport

Events and their arguments form graphs on both the text and image sides. The text event graph G_t has nodes for the primary event and for each argument role with its entity. The image event graph G_i has nodes for the event and for each detected object. The goal is to align these two graphs so that event nodes and argument nodes match across modalities.

A distance between the two graphs is defined using optimal transport:

$$d(G_t, G_i) = \min_{\mathbf{T}} \mathbf{T} \odot \mathbf{C}. \tag{4}$$

Here $\mathbf{C}$ is a cost matrix whose entry C_{ab} measures the distance between node a in the text graph and node b in the image graph. The matrix $\mathbf{T}$ is a transport plan of size $n \times m$, where n and m are the numbers of nodes in G_t and G_i respectively. The notation $\odot$ denotes the Hadamard (element-wise) product, and the minimization effectively searches for the best way to transport probability mass from nodes in the text graph to nodes in the image graph under the cost matrix $\mathbf{C}$. Intuitively, equation (4) produces a global alignment score between two graphs that takes all node-to-node correspondences into account.

Cost between event nodes. The cost between a text event node v and the image is defined as

$$C(v, i) = c(\mathbf{t}_v, \mathbf{i}) + c(\boldsymbol{\phi}_v, \mathbf{i}). \tag{5}$$

The vector $\mathbf{t}_v$ is the contextualized embedding of the event trigger word in the caption, and $\boldsymbol{\phi}_v$ represents the event type embedding (such as TRANSPORT). This cost is small when both the event word and event type are compatible with the image.

Cost between argument nodes. For each argument consisting of a role a and an entity e, and for each image object o, the cost is

$$C(\langle a, e\rangle, o) = d(a, o) + d(e, o) = c(\mathbf{t}_a, \mathbf{i}_o) + c(\mathbf{t}_e, \mathbf{i}_o) + c(\boldsymbol{\phi}_e, \boldsymbol{\phi}_o). \tag{6}$$

The vector $\mathbf{t}_a$ encodes the textual description of the argument role, such as "the entity being transported", while $\mathbf{t}_e$ and $\mathbf{i}_o$ are the entity and object embeddings as in equation (3). This formulation forces both the role description and the entity mention to align with the same visual object, and also encourages type consistency through the third term.

Sinkhorn-based computation of the transport plan. Directly solving the optimal transport problem in equation (4) is computationally expensive. CLIP-Event uses an entropically regularized version solved by the Sinkhorn–Knopp algorithm. The closed-form expression for the transport plan is

$$\mathbf{T} = \mathrm{diag}(\mathbf{p})\ \exp(-\mathbf{C}/\gamma)\ \mathrm{diag}(\mathbf{q}), \tag{7}$$

where $\mathbf{p} \in \mathbb{R}_+^n$ and $\mathbf{q} \in \mathbb{R}_+^m$ are scaling vectors that enforce marginal constraints, $\exp$ is applied element-wise, γ is a temperature parameter controlling the smoothness of the transport, and $\mathrm{diag}(\cdot)$ forms a diagonal matrix from a vector.

The scaling vectors are obtained by iterating

$$\mathbf{p}^{(k+1)} = \mathbf{1} \oslash (\mathbf{K}\mathbf{q}^{(k)}), \qquad \mathbf{q}^{(k+1)} = \mathbf{1} \oslash (\mathbf{K}^\top \mathbf{p}^{(k+1)}), \tag{8}$$

where $\mathbf{K} = \exp(-\mathbf{C}/\gamma)$, $\oslash$ denotes element-wise division and $\mathbf{1}$ is an all-ones vector. After a finite number of iterations k, the approximate transport plan is

$$\mathbf{T}^{(k)} = \mathrm{diag}(\mathbf{p}^{(k)})\, \mathbf{K}\, \mathrm{diag}(\mathbf{q}^{(k)}). \tag{9}$$

This procedure is differentiable, which allows gradients to flow through the graph alignment process during training. Intuitively, the Sinkhorn updates repeatedly normalize rows and columns of the transport matrix so that the final plan distributes probability mass across alignments of text nodes and image nodes in a balanced manner.

5.5 Event Structure Driven Negative Sampling

To force the encoders to pay attention to event types and argument roles, CLIP-Event designs negative samples that are hard and structure-aware.

Negative event sampling. The authors first run a CLIP-based event type classifier on the pretraining data and build a confusion matrix between predicted event types and gold primary event types. For each image, event types that are often confused with the primary event type are treated as negative but challenging alternatives. This captures visually similar but semantically distinct events, such as *transport* versus *arrest*.

Negative argument sampling. Argument roles for an event are ordered according to the ontology, such as *AGENT*, *ENTITY*, *INSTRUMENT*. Negative arguments are created by rotating this sequence, assigning entities to incorrect roles. For instance, an *injured man* that should be an *ENTITY* may be turned into an *AGENT*. When an event has only one argument, negative roles are sampled according to the confusion matrix of a text argument extraction system.

Prompt-based description generation. The model transforms event structures into textual descriptions using a set of prompt functions:

- **Single-template prompt**: encodes all arguments into one sentence by filling placeholders.
- **Composed-template prompt**: decomposes arguments into multiple simple sentences, one per role.
- **Continuous prompt**: inserts learnable tokens into the text sequence to encode structural markers.
- **Caption editing**: minimally edits original captions by changing the event trigger or swapping roles, which preserves naturalness.
- **GPT-3-based prompt**: uses a large language model conditioned on example event structures to generate fluent positive and negative descriptions for each event.

These descriptions are used both as positive examples (correct event structures) and as negative examples (corrupted event types or roles) in contrastive learning.

5.6 Contrastive Learning Objective

The training objective combines instance-level contrastive learning on descriptions and image–text pairs with graph-level alignment via optimal transport.

Description-level contrastive loss. For each image–description pair $\langle t, i \rangle$, a Kullback–Leibler divergence loss encourages positive descriptions to have similarity close to one and negative descriptions to have similarity close to zero:

$$L_1 = \sum_{\langle t,i \rangle} D_{\mathrm{KL}}\big(s(t,i) \,\|\, \mathbf{1}_{\{t \in T^+\}}\big). \tag{10}$$

Here $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$ is the Kullback–Leibler divergence between two Bernoulli distributions. The indicator $\mathbf{1}_{\{t \in T^+\}}$ equals one if t is a positive description for image i and zero otherwise. This loss function treats the similarity $s(t,i)$ as a predicted probability and penalizes deviations from the desired label. Descriptions of other images within the same batch automatically act as additional negatives.

Event graph alignment loss. To integrate graph-level alignment, the model also minimizes the distance between text event graphs and image event graphs:

$$L_2 = \sum_{\langle t,i \rangle} d(G_t, G_i). \tag{11}$$

The distance $d(G_t, G_i)$ is computed using the optimal transport formulation in equations (4)–(9). Minimizing L_2 encourages consistent alignment between events and arguments across modalities, not just between entire images and captions.

Joint objective. The final training loss is a weighted sum of the two components:

$$L = \lambda_1 L_1 + \lambda_2 L_2. \tag{12}$$

The weights λ_1 and λ_2 balance the importance of instance-level and graph-level objectives and are set to one in the experiments. Intuitively, equation (12) teaches the model both to distinguish correct versus corrupted descriptions and to respect detailed event-argument alignments.

5.7 Model Architecture and Hyperparameters

The encoders are initialized with the public CLIP model using the ViT-B/32 vision backbone and the corresponding text Transformer. The pretraining uses image–caption pairs from VOANews and the generated positive and negative event descriptions. The Sinkhorn iterations are run for a finite number of steps to keep the graph alignment computationally feasible. Detailed choices of learning rate, batch size and optimization schedule follow standard CLIP practice and are reported in the original appendix.

6 Empirical Results

6.1 Event Extraction on M2E2 and SWiG

Table 4 of the original paper reports detailed results for multimedia event extraction and grounded situation recognition. The main observations are as follows.

Zero-shot event extraction. On the M2E2 dataset in zero-shot settings:

- A baseline CLIP model achieves an event F1-score of 40.7 and an argument F1-score of 10.7.
- CLIP pretrained on the VOANews news data improves event F1 to 42.6 and argument F1 to 11.1, indicating that domain-matched pretraining alone provides modest gains.
- CLIP-Event further increases event F1 to 48.1 and argument F1 to 14.8. This corresponds to an absolute improvement of 5.5 points in event F1 and a 33.3 percent relative improvement in argument F1 over the CLIP baseline.

These results show that structured event supervision significantly enhances zero-shot capability for both event recognition and argument extraction.

Effect of optimal transport and prompts. The ablation variant *w/o OptimalTransport*, which removes graph alignment and trains only with L_1, yields an event F1 of 44.1 and argument F1 of 11.9. Both scores are clearly below those of full CLIP-Event, demonstrating that optimal transport based graph alignment contributes substantially, especially to argument extraction.

Variants that use a single prompt function during training (Single Template, Composed Template, Continuous Prompt, Caption Editing, GPT-3 Prompt) perform between the CLIP pretrained-on-news baseline and the full CLIP-Event model. This indicates that combining multiple prompt strategies and graph alignment yields the strongest improvements.

On SWiG for grounded situation recognition, CLIP-Event improves several F1-style metrics over CLIP and CLIP pretrained on news, again confirming the benefit of event-centric supervision.

Supervised settings. When fine-tuned on SWiG, CLIP-Event combined with SWiG supervision achieves an event F1 of 52.7 on M2E2, outperforming both the state-of-the-art supervised model and CLIP fine-tuned on SWiG. The ablation without optimal transport lags behind, reinforcing the advantage of graph alignment even with direct supervision.

6.2 Image Retrieval on Flickr30k, MSCOCO and VOANews

Table 5 reports Recall@1 for text-to-image and image-to-text retrieval. CLIP-Event consistently improves over the CLIP baseline and the news-pretrained CLIP across all datasets.

For example:

- On VOANews text-to-image retrieval, CLIP achieves a Recall@1 of 21.2. CLIP pretrained on news reaches 23.5, while CLIP-Event attains 27.5, a substantial improvement on this challenging dataset.
- On Flickr30k and MSCOCO, CLIP-Event also increases Recall@1 by several points for both retrieval directions. The gains are smaller than on VOANews, which reflects that VOANews is more event-centric and benefits more from structured event alignment.

The variant without optimal transport performs better than CLIP but worse than full CLIP-Event, confirming that event graph alignment improves fine-grained retrieval, especially when multiple similar objects are present.

6.3 Visual Commonsense Reasoning and VisualCOMET

Table 6 evaluates zero-shot performance on VCR and VisualCOMET.

- On VCR, CLIP reaches an answer F1 of 51.1 and rationale F1 of 46.8. CLIP-Event improves these to 52.4 and 49.2, respectively. The larger gain in rationale F1 suggests that detailed event knowledge helps select rationales that refer to specific interactions and roles in the scene.
- On VisualCOMET, which measures Accuracy@50 for predicting plausible intents, CLIP obtains 20.1, while CLIP-Event achieves 22.4. The improvement indicates that event-centric pretraining aids temporal commonsense reasoning about what participants intend to do.

The ablation without optimal transport again lies between CLIP and CLIP-Event, indicating that graph alignment contributes across different reasoning tasks.

6.4 Overall Findings

Across all tasks, CLIP-Event exhibits the following trends:

- Substantial gains in zero-shot event extraction and argument localization.
- Consistent improvements in image retrieval, especially on the event-rich VOANews dataset.
- Noticeable benefits for visual and temporal commonsense reasoning benchmarks that rely on understanding who did what to whom and why.

These results collectively validate that injecting event structures and graph alignment into vision–language pretraining yields more robust and semantically rich representations.

7 Summary

7.1 Research Goal and Contributions

The central goal of CLIP-Event is to extend vision–language pretraining so that models understand not just which objects appear in an image, but also which events occur and how entities participate in those events. The study makes several key contributions:

- Construction of an event-rich VOANews dataset with automatically extracted event structures from long news captions.
- A contrastive pretraining framework that uses multiple prompt-based event descriptions and hard negative sampling to emphasize event types and argument roles.
- An event graph alignment module based on optimal transport, which aligns text event graphs and image event graphs at node and structure levels.
- Extensive experiments showing that the resulting model improves over CLIP and other baselines on multimedia event extraction, grounded situation recognition, image retrieval, visual commonsense reasoning and temporal event reasoning.

7.2 Innovation Compared with Previous Methods

Relative to previous CLIP-style models and event extraction systems, CLIP-Event introduces several innovations:

- **Structured event supervision during pretraining**: Rather than relying only on global image–caption pairs, the model explicitly leverages event types, argument roles and entity mentions extracted from text.
- **Event-aware negative sampling and prompts**: The model generates challenging negative descriptions by manipulating event types and arguments and uses a diverse set of prompt functions, including GPT-3-generated descriptions, to cover different styles of language.
- **Cross-modal event graph alignment**: The optimal transport based alignment between text and image event graphs encourages fine-grained correspondences that go beyond object labels, capturing who plays which role in which event.
- **Unified pretraining for multiple downstream tasks**: The same pretrained model, with or without light adaptation, supports event extraction, retrieval and commonsense reasoning tasks, illustrating the generality of the learned event-centric representations.

7.3 Limitations and Future Directions

Despite its strengths, CLIP-Event has several limitations:

- **Dependence on external information extraction**: The quality of event and argument annotations depends on the accuracy of the text information extraction system and the underlying ontology. Errors or coverage gaps in this pipeline can propagate into pretraining.
- **Domain bias toward news imagery**: VOANews focuses on news photos, which may limit generalization to other domains such as everyday social media images or specialized scientific imagery.
- **Computational cost of optimal transport**: Sinkhorn-based optimal transport over event graphs introduces additional computational overhead compared with standard CLIP training, which may restrict scalability to larger graphs or higher-resolution images.
- **Fixed event ontology**: The model relies on a predefined ontology with a fixed set of event types and roles. Extending to new event schemas or more open-ended event descriptions requires additional engineering.

Future research directions include developing more robust and domain-adaptive event extraction pipelines, designing lighter-weight or approximate graph alignment mechanisms, and exploring open-vocabulary or schema-free representations of events that can generalize beyond a fixed ontology.

7.4 Key Takeaways

From a high-level perspective, three main lessons emerge:

1. **Event structures matter**: Explicitly modeling event types and argument roles during vision–language pretraining leads to substantial gains in both structured prediction and high-level reasoning tasks.

2. **Graph alignment is powerful**: Optimal transport based alignment between text and image event graphs provides a principled way to connect multi-node structures across modalities, which improves fine-grained localization and retrieval.

3. **Structured supervision complements large-scale contrastive learning**: Even when starting from a strong CLIP baseline, adding structured event knowledge and carefully designed prompts yields consistent improvements across diverse benchmarks.