A. Implementation Details

A.1. Event Knowledge Extraction Details.

Text Knowledge Extraction Details. We use the state-of-the-art text information extraction tools OneIE [5]. In detail, we run the dockerized version GAIA [3] that is using the DARPA AIDA event ontology [1], the most fine-grained text event ontology, attached in *event_ontology_oneie.json*.

Example Event Types	Arguments
ArtifactExistence.ArtifactFailure.MechanicalFailure	MechanicalArtifact, Instrument, Place
ArtifactExistence.DamageDestroy.Damage	Damager, Artifact, Instrument, Place
ArtifactExistence.DamageDestroy.Destroy	Destroyer, Artifact, Instrument, Place
ArtifactExistence.Shortage.Shortage	Experiencer, Supply, Place
Conflict.Attack	Attacker, Target, Instrument, Place
Conflict.Attack.AirstrikeMissileStrike	Attacker, Target, Instrument, Place
Conflict.Attack.BiologicalChemicalPoisonAttack	Attacker, Target, Instrument, Place
Conflict.Attack.Bombing	Attacker, Target, Instrument, Place
Conflict.Attack.FirearmAttack	Attacker, Target, Instrument, Place
Conflict.Attack.Hanging	Attacker, Target, Instrument, Place
Conflict.Attack.Invade	Attacker, Target, Instrument, Place
Conflict.Attack.SelfDirectedBattle	Attacker, Target, Instrument, Place
Conflict.Attack.SetFire	Attacker, Target, Instrument, Place
Conflict.Attack.Stabbing	Attacker, Target, Instrument, Place
Conflict.Attack.StealRobHijack	Attacker, Target, Instrument, Place
Conflict.Attack.Strangling	Attacker, Target, Instrument, Place
Disaster.AccidentCrash.AccidentCrash	DriverPassenger, Vehicle, CrashObject, Place
Disaster.DiseaseOutbreak.DiseaseOutbreak	Disease, Victim, Place
Disaster.FireExplosion.FireExplosion	FireExplosionObject, Instrument, Place
Justice.ArrestJailDetain.ArrestJailDetain	Jailer, Detainee, Crime, Place
Justice.InitiateJudicialProcess	Prosecutor, Defendant, JudgeCourt, Crime
Justice.InitiateJudicialProcess.ChargeIndict	Prosecutor, Defendant, JudgeCourt, Crime
Justice.InitiateJudicialProcess.TrialHearing	Prosecutor, Defendant, JudgeCourt, Crime
Justice.Investigate	Investigator, Defendant, Place
...	...

Table 1. Example event types from Text Information Extraction system, the full list is attached in *event_ontology_oneie.json*.

In addition, we explore open-world event extraction that is not limited to a specific event ontology. We apply OpenIE tools [1, 10], which output $\langle$*subject*, *relation*, *object*$\rangle$. For example, from the caption in Fig. 2 in the main paper, OpenIE extracts $\langle$*protesters*, CARRY, *injured man*$\rangle$, $\langle$*clashes*, WITH, *riot police*$\rangle$, and $\langle$*Independence Square*, IN, *Kyiv*$\rangle$. However, from 100 randomly selected captions, we find that 72.1% events from OpenIE are not visually detectable, such as THINKING and INVITING. Considering that these events will introduce a lot of noise to the cross-media alignment, we only adopt the aforementioned supervised IE model to obtain event knowledge from text.

Visual Knowledge Extraction Details. We apply Faster R-CNN [9] to detect objects, which is trained on Open Images [2] with 600 object types (classes). For event knowledge extraction on images, the most similar tool is grounded situation recognition [8], which achieves 39.6% accuracy on event extraction. Considering the errors propagated from extraction models, instead of extracting event knowledge from images as a supervision signal, we take advantage of text information extraction that have better event extraction performance (75.2% on F-score), to provide supervision to enhance visual event understanding.

[1] https://github.com/NextCenturyCorporation/AIDA-Interchange-Format/blob/master/java/src/main/resources/com/ncc/aif/ontologies/LDCOntologyM36

A.2. Parameter Settings

We utilize the Text and Vision Transformers of "ViT-B/32" to initialize our encoders. The batch size is 128. We set the learning rate as $1e-6$ with a linearly-decaying schedule. We train 20 epochs with Adam [6] as the optimizer, and select the best model based on the image-retrieval performance on VOANews testing dataset. The optimal transport plan is obtained within $k = 50$ iterations. To get the bounding box embeddings from CLIP visual backbone, we extract grid features and perform average pooling on the grids covered by the bounding box. For CLIP-ViT-B models, we reshape the patch representation of the final layer into grid features. For CLIP-ResNet models, we use the grid features from the last layer before the pooling. The model is trained on eight Tesla V100 GPUs with 32GB DRAM, and the pretraining takes around one day.

A.3. Multimedia Event Extraction Implementation Details

Task Setting. Multimedia Event Extraction [4] aims to (1) classify images into eight event types, and (2) localize argument roles as bounding boxes in images.

Evaluation Goal. We choose this task as a direct assessment of event structure understanding.

Our Approach. Under zero-shot settings, we directly evaluate the pretraining model on the testing set. We evaluate the event extraction and argument extraction on all images, which contain visual events of 8 types. We add OTHER to detect the images not belonging to the eight target types. The description of OTHER is *An image of other events.* For argument extraction, we rank argument roles for each object bounding box, and also add OTHER argument role as a candidate with the description *other roles of the event*.

Under supervised settings, we use the same training data SWiG as the sate-of-the-art model [4], but replacing the text event table with the annotation table, and setting the optimal transport plan as the fine-grained alignment between event graphs. We use the same training dataset SWiG [8] with 125k images to further finetune our model to compare with the supervised models. During finetuning, we replace the text event extraction results with the annotated events for images, and set the optimal transport plan as the ground truth alignment between arguments and object bounding boxes.

A.4. Grounded Situation Recognition Implementation Details

Task Setting. Grounded Situation Recognition [8] selects an event type from 504 verbs, and predicts the entity name and the bounding box for each argument role.

Evaluation Goal. It is also a direct evaluation of event structure understanding, but with larger size of event types and argument roles.

Implementation. Grounded Situation Recognition requires the model to assign each image to a verb from 504 verbs (such as RIDING), and name the argument (such as *man*) of each argument role (such as AGENT). For each image, we rank the verbs using the description "*An image of* $\langle verb \rangle$". For each argument role, we obtain the candidate names from the training set, and rank the candidate names using the description "*The* $\langle name \rangle$ *is a* $\langle role \rangle$ *of* $\langle verb \rangle$", such as "*The man is a agent of riding*". For each object, we rank argument roles including OTHER, similarly to Multimedia Event Extraction. Following [8], we ignore the PLACE argument role since it always not appear in the images. The supervised setting is the same as Multimedia Event Extraction.

Evaluation Metrics. We follow [8] to evaluate the accuracy of verb prediction (*verb*), argument name prediction (*value* for each argument and *value-all* for all arguments of an event), and argument bounding box and name prediction (*ground* for each argument and *ground-all* for all arguments).

A.5. VCR Implementation Details

Task Setting. VCR is a question answering task[2], including (1) *Answer Prediction* from four options, and (2) *Rationale Prediction* from four options to support the aforementioned answer.

Evaluation Goal. We include this task to evaluate whether event understanding can better support downstream tasks. To evaluate the quality of pretraining models, we adopt zero-shot settings solely relying on image-text alignment for a fair comparison.

Implementation. For Answer Prediction, we rank answers concatenated with questions. For Rationale Prediction, we rank rationales by concatenating the question, the answer and the rationale. The ranking is based on both image alignment $d(i, t)$ and event graph alignment $d(G_i, G_t)$. We also consider the question as query and concatenate them with the answer during ranking.

A.6. VisualCOMET Implementation Details

Task Setting. Given the image and the event happening in the image with its participants, VisualCOMET [7] aims to generate "intents" showing what the participants "*need to do*" before the image event, "*want to do*" during the image event, and "*will most likely to do*" after the image event.

Goal. It necessitates a deep grasp of events and their connections, as well as a thorough comprehension of arguments roles.

Implementation. The input of VisualCOMET[3] is an image with events and participants, as shown in Fig. 1. The output are intents, which is a short description of an event, such as "*swim to safety*", "*sink in the water*", etc. For each image and participant, we use intents from the training data as candidate intents, and rank them based on both image alignment $d(i, t)$ and event graph alignment $d(G_i, G_t)$. The text is the concatenation of (1) input event description, (2) a temporal description (including "*before person1 need to*", "*because person1 need to*" and "*after person1 will most likely to*"), and (3) the candidate intents. For example, given the image with the input event "*person1 is trying to escape from the water*", we concatenate it with the temporal description "*because person1 wanted to*" and the candidate intent "*swim to safety while*". The ranking is based on both image alignment $d(i, t)$ and event graph alignment $d(G_i, G_t)$, similar to Visual Commonsense Reasoning.

[2] https://visualcommonsense.com/

[3] https://visualcomet.xyz/

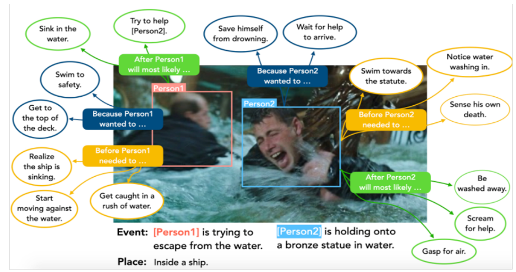

Figure 1. An example from VisualCOMET [7].

B. Effect of Text Information Extraction Performance

Since text information extraction may have errors, we analyze its performance in the following sections.

B.1. Text Event Extraction Performance Table

The extraction performance of each component is shown in Tab. 2, which achieves 72.1% F-score on event extraction.

Component		**Benchmark**	**Metric**	**Score**
Event Mention Extraction	Entity	ACE+ERE	F_1	90.2
	Trigger	ACE+ERE	F_1	72.8
	Argument	ACE+ERE	F_1	54.8
	Relation	ACE+ERE	F_1	49.5
Document-level Argument Extraction		ACE	F_1	66.7
		RAMS	F_1	48.6
Coreference Resolution	Entity	OntoNotes	CoNLL	92.4
	Event	ACE	CoNLL	84.8
	Event	ERE-ES	CoNLL	81.0

Table 2. Performance (%) of each component.

B.2. Event Type distribution

As shown in Fig. 2, the events extracted from captions are primarily visually detectable events, i.e., the. events can

Model	Flickr30k						MSCOCO						VOANews					
	text-to-image			image-to-text			text-to-image			image-to-text			text-to-image			image-to-text		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	62.2	85.9	91.7	81.9	95.0	97.5	30.3	55.0	66.4	50.3	75.7	84.0	21.2	63.4	74.7	23.4	63.1	73.9
CLIP pretrained on news	64.3	87.5	92.7	81.2	95.4	98.2	32.2	57.4	68.4	50.8	75.6	83.8	23.5	69.5	79.9	25.1	70.2	80.1
CLIP-Event	**67.0**	**89.0**	**93.9**	**82.6**	**95.9**	**98.4**	**34.0**	**59.4**	**70.5**	**51.3**	**76.0**	**84.0**	**27.5**	**70.7**	**82.1**	**28.7**	**71.0**	**81.0**
w/o OptimalTransport	65.6	88.3	93.6	80.5	94.8	97.4	32.5	58.0	68.9	51.0	75.2	82.9	25.5	70.6	80.7	26.9	70.4	80.5

Table 3. R@1(%), R@5(%), R@10(%) on image retrieval on Flickr30k (1k test), MSCOCO (5k test) and VOANews.

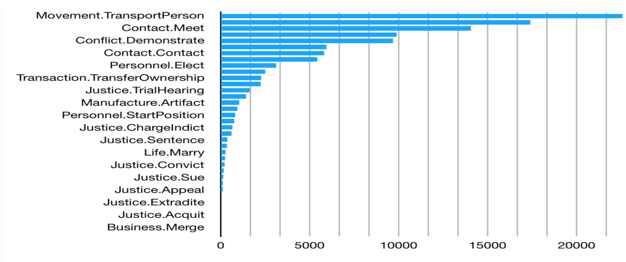

Figure 2. The top frequent event types from the event extraction results on VOANews captions.

be depicted in the images.

References

[1] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, 2015. 1

[2] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1

[3] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, et al. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, 2020. 1

[4] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, 2020. 1

[5] Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, 2020. 1

[6] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 1

[7] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer, 2020. 2

[8] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer, 2020. 1, 2

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[10] Michael Schmitz, Stephen Soderland, Robert Bart, Oren Etzioni, et al. Open language learning for information extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 523–534, 2012. 1