Multimodal Cross-Document Event Coreference Resolution Using Linear Semantic Transfer and Mixed-Modality Ensembles

Abhijnan Nath[1], Huma Jamil[1], Shafiuddin Rehan Ahmed[2], George Baker[2],
Rahul Ghosh[1,3*], James H. Martin[2], Nathaniel Blanchard[1], and Nikhil Krishnaswamy[1]
[1]Colorado State University, Fort Collins, CO, USA
[2]University of Colorado, Boulder, CO, USA
[3]Purdue University, West Lafayette, IN, USA
{abhijnan.nath, nkrishna}@colostate.edu

Abstract

Event coreference resolution (ECR) is the task of determining which distinct mentions of events within a multi-document corpus are actually linked to the same underlying occurrence. Images of the events can help facilitate resolution when language is ambiguous. Here, we propose a multimodal cross-document event coreference resolution method that integrates visual and textual cues with a simple linear map between vision and language models. As existing ECR benchmark datasets rarely provide images for all event mentions, we augment the popular ECB+ dataset with event-centric images scraped from the internet and generated using image diffusion models. We establish three methods that incorporate images and text for coreference: 1) a standard fused model with finetuning, 2) a novel linear mapping method without finetuning and 3) an ensembling approach based on splitting mention pairs by semantic and discourse-level difficulty. We evaluate on 2 datasets: the augmented ECB+, and AIDA Phase 1. Our ensemble systems using cross-modal linear mapping establish an upper limit (91.9 CoNLL F1) on ECB+ ECR performance given the preprocessing assumptions used, and establish a novel baseline on AIDA Phase 1. Our results demonstrate the utility of multimodal information in ECR for certain challenging coreference problems, and highlight a need for more multimodal resources in the coreference resolution space.

Keywords: event coreference resolution, multimodality, ensemble methods

1. Introduction

Imagine two newspaper articles about the same event. The articles come from different sources with radically different perspectives and report the event with very different language. They use different action verbs, include ambiguous pronominal references, describe causes differently, and even attribute different intentionality to the event—for example, "*Buzina, 45, was shot dead*" vs. "*He was murdered*". An automated system may be unable to identify from the text alone that the two events described are actually the same. This is the problem of *cross-document coreference resolution* (CDCR) of events: inferring that two event mentions in different documents actually refer to the same thing.

Now imagine that each of the articles is accompanied by an image. While not identical, they clearly contain the same people, entities, and actions. This would be strong evidence to a reader that the two events described in the different articles are in fact the same.

Purely text-based approaches to CDCR, while built on sophisticated Transformer-based language models (LMs) (Vaswani et al., 2017; Beltagy et al., 2020), are blind to such potentially useful multimodal information. This problem is exacerbated by the relative dearth of multimodal information included in event CDCR corpora.

In this work, we propose a novel multimodal *event* CDCR method. Where current state-of-the-art coreference approaches that consider visual information demonstrate the utility of a multimodal approach, they do so at a high computational cost (Guo et al., 2022). Furthermore, they typically focus on linking objects rather than events. We address the sparsity of multimodal data in benchmark datasets by retrieving images associated with the metadata of event mentions, and generating event-centric images with state-of-the-art image diffusion models. We perform coreference experiments in a fully multimodal setting and rigorously test the contribution of multimodal information to CDCR.[1]

In total, our novel contributions include:

- A novel approach to multimodal *cross document event coreference* (MM-CDCR) including a low-compute, bidirectional *linear semantic transfer* technique (`Lin-Sem`) based on *semantic equivalence* across modalities;
- A model ensemble hybrid approach that applies text-only or multimodal methods to different categories of mention pairs based on their semantic and discourse-level difficulty;
- A novel method for enriching text-only coreference datasets (e.g., ECB+ (Cybulska and

[*]This work conducted at Colorado State University.

[1]Our code can be accessed at https://github.com/csu-signal/multimodal-coreference.

Vossen, 2014)) with event-centric images using generative image diffusion;
- A new benchmark result on the AIDA Phase 1 dataset (Tracey et al., 2022), an explicitly multimodal event CDCR dataset. To our knowledge, this is the first evaluation performed over this dataset.

2. Related Work

Cross-Document Event Coreference Resolution Most previous works on CDCR have been limited to *text-only* (Eisenstein and Davis, 2006; Chen et al., 2011). Early works (e.g., Humphreys et al. (1997); Bagga and Baldwin (1999); Chen and Ji (2009)) used supervised training over features like part-of-speech tags, phrasal-matching, or aligned arguments. While Kenyon-Dean et al. (2018) enhanced lexical features with "static" embeddings like contextual word2vec (Mikolov et al., 2013), most recent works (Yu et al., 2022; Caciularu et al., 2021; Yadav et al., 2021) uses latent representations from Transformer-based encoders to compute pairwise mention scores of possible antecedents. Works such as Held et al. (2021) and Ahmed et al. (2023) overcome the quadratic complexity of the mention pair architecture by pruning negative pairs using discourse-coherence and lexical similarity (synonymous lemma pairs) respectively. We use Ahmed et al. (2023)'s "oracle" assumption for our pruning procedure.

Multimodal Frameworks Most previous works in multimodal vision-language processing (e.g., (Le et al., 2019; Tan and Bansal, 2019)) have been compute-intensive, using separate encoders for visual and linguistic inputs, and auxiliary encoders for cross-modal or query-related modeling. High-performing but high-compute models like ViL-BERT (Lu et al., 2019) concatenate embeddings from different modalities before fine-tuning. Works such as Li et al. (2020), Tong et al. (2020), and Chen et al. (2021) leverage a common representation space for coreference-adjacent tasks like event extraction and detection in images and videos, but emphasize finding relations within a document or a topic. Works specific to multi-modal *entity* coreference resolution such as Guo et al. (2022) treat it largely as a grounding problem, using graph networks to link references in dialogue to items in a scene before feeding representations into BERT-style encoders to resolve scene-based visual-linguistic coreference chains. Our work is multimodal, cross-document, *and* event focused, and performs faster with the aid of linear mappings.

Linear Projection Across Neural Networks Previous research within computer vision has explored using affine (McNeely-White et al., 2020, 2022; Jamil et al., 2023) as well as non-linear (Lenc and Vedaldi, 2015) transformations to explore equivalence of *unimodal* function approximators like CNNs. They show that two distinct, highly non-linear neural networks can learn similar properties transferable up to a linear projection while retaining near-equivalent performance on tasks like image classification or facial recognition. Similar techniques using affine mappings were reported in Merullo et al. (2023), who explore the equivalence of such approximators *across* modalities while also casting new light on high-fidelity transfer of non-linguistic features into a generative LLM via unidirectional linear projections from image spaces. Nath et al. (2022) demonstrated linear mappings also preserve information across language models. Ghaffari and Krishnaswamy (2023) showed the same between language models and neural networks trained over tabular data.

We use a low-compute, cross-modal, *bidirectional* linear-mapping technique (`Lin-Sem`: **Lin**ear **Sem**antic Transfer) between language and vision Transformers, on the challenging event coreference task. We demonstrate where this linear transfer is providing useful information toward coreference resolution compared to a text-only discriminative LLM, or fused modality models following standard fine-tuning.

3. Methodology

Fig. 1 illustrates the pipeline for our methodology, the components of which are detailed as follows.

Semantic Equivalence

$$\mathcal{V}\big(x,y,\phi(x,y)\big):R^{n\times w\times h\times 3}\rightarrow R^{n\times H} \quad (1)$$

$$\mathcal{LLM}\big(x,y,\phi(x,y)\big):R^{n\times m}\rightarrow R^{n\times H} \quad (2)$$

Let (1) and (2) represent the heterogeneous image and text representations for vision and text Transformer models respectively. $(x, y) \in \chi$ represents all the pairs of samples in sample space χ, $\phi(x, y)$ represents the concatenation of the image or text pair in their respective modalities, n and H represent the total sample pairs and hidden dimensions respectively, and m is the LLM's max token-length.

We define cross-modal semantic equivalence as follows: two representations $\mathcal{V}$ and $\mathcal{LLM}$ in distinct modalities are semantically equivalent if there exists a bidirectional map $M_{\mathcal{V}\leftrightarrow\mathcal{LLM}}$ s.t.:

$$\forall_{x,y\in\chi}:\mathcal{V}\big(x,y,\phi(x,y)\big)\approx M_{\mathcal{LLM}\rightarrow\mathcal{V}}\,\mathcal{LLM}\big(x,y,\phi(x,y)\big) \quad (3)$$

$$\forall_{x,y\in\chi}:\mathcal{LLM}\big(x,y,\phi(x,y)\big)\approx M_{\mathcal{V}\rightarrow\mathcal{LLM}}\,V\big(x,y,\phi(x,y)\big) \quad (4)$$

while assuming both $\mathcal{V}$ and $\mathcal{LLM}$ to be bijective or invertible, so,

$$M_{\mathcal{LLM}\rightarrow\mathcal{V}}=\mathcal{V}\big(x,y,\phi(x,y)\big)\circ\mathcal{LLM}\big(x,y,\phi(x,y)\big)^{-1} \quad (5)$$

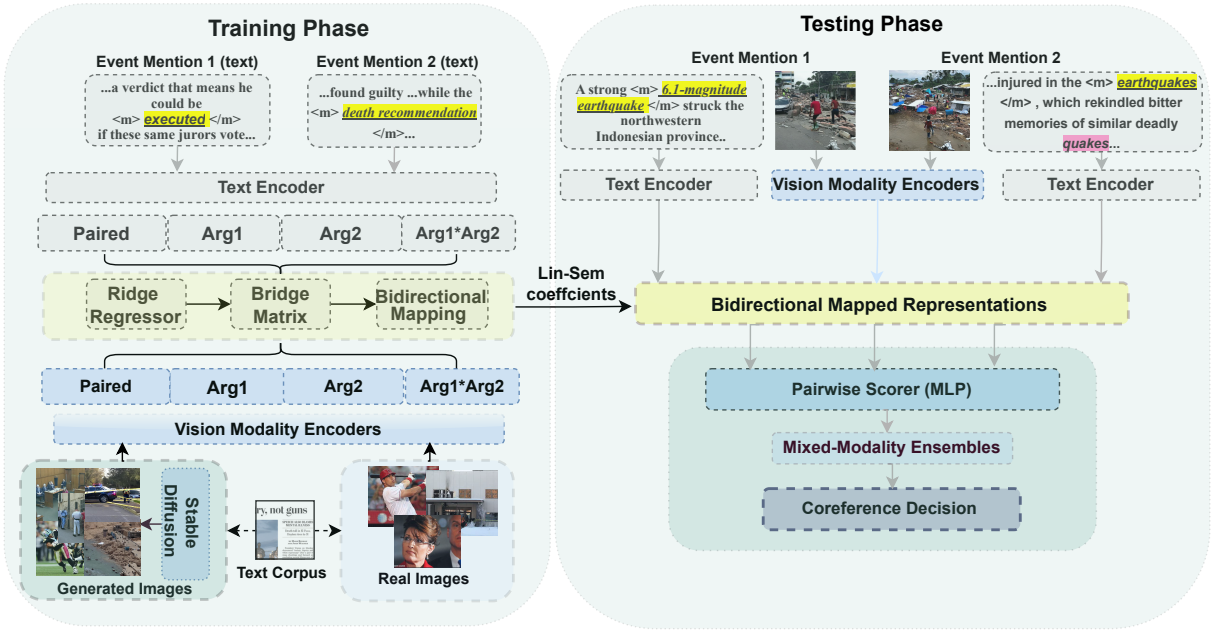


Figure 1: Our approach for Multimodal CDCR using `Lin-Sem`. Linear Mapping (`Lin-Sem`) procedure between the distinct text and image embedding spaces for an event pair in the ECB+ corpus. `Arg1` and `Arg2` refer to the individual images in the pair and the trigger events (in yellow) surrounded by the `<m>` and `</m>` special tokens embedded in the text-encoder (LLM).

$$M_{\mathcal{V}\rightarrow\mathcal{LLM}} = \mathcal{LLM}\left(x,y,\phi(x,y)\right) \circ \mathcal{V}\left(x,y,\phi(x,y)\right)^{-1} \quad (6)$$

Since a closed-form solution to analytically derive the mapping function $M_{\mathcal{V}\leftrightarrow\mathcal{LLM}}$ is not always feasible and since many task-based fine-tuning heads over a Transformer-based LLM involve fitting a linear classification layer, we propose a parameter-efficient linear-mapping technique `Lin-Sem`. We estimate the mapping function within a empirical risk minimization framework by using a ridge regression between the two cross-modal representations. Mathematically,

$$M_{\mathcal{LLM}\rightarrow\mathcal{V}} \leftarrow \text{minimize } ((\mathcal{V}-\beta\mathcal{LLM})^T(\mathcal{V}-\beta\mathcal{LLM})+\lambda\beta^T\beta) \quad (7)$$

$$M_{\mathcal{V}\rightarrow\mathcal{LLM}} \leftarrow \text{minimize } ((\mathcal{LLM}-\beta\mathcal{V})^T(\mathcal{LLM}-\beta\mathcal{V})+\lambda\beta^T\beta) \quad (8)$$

We assume λ=1 while β represents the L^2-norm regularization parameter.

Datasets We evaluated our methods on the ECB+ (Cybulska and Vossen, 2014) and the AIDA Phase 1 (Tracey et al., 2022) datasets. While the former is a popular, English-only CDCR benchmark containing a diverse range of news articles, the latter contains multimodal resources specific to Russia-Ukraine relations, in English, Russian, and Ukrainian. We focus only on the English documents.[2] For our experiments, we used training and evaluation splits following Cybulska and Vossen (2015) for ECB+ and Tracey et al. (2022) for AIDA Phase 1. Table 1 shows corpus-level statistics for these two datasets.

	ECB+			AIDA Phase 1	
Split	Train	Dev	Test	Practice	Eval
Docs	594	196	206	63	69
Event Mentions	3808	1245	1780	603	846
Clusters	1464	409	805	186	270
Singletons	1053	280	623	132	197
Images	3808*	1245*	1780*	417	662

Table 1: ECB+ and AIDA corpus-level statistics. Tracey et al. (2022) refers to the provided train and test sets as "practice" and "eval", respectively.

*Including images generated using Stable Diffusion.

Augmenting ECB+ with Images Since ECB+ does not provide images in their metadata, we scraped through the links provided in the documents and searched the Internet Archive for archived versions of articles with dead links. For original ECB documents without links, we manually search for keywords to retrieve articles. Out of 502 ECB+ document links, 43% were broken, but 50% could be recovered us-

[2]The AIDA Phase 1 dataset was created for the DARPA Active Interpretation of Disparate Alternatives (AIDA) program and is available from the Linguistic Data Consortium (catalog number LDC2019E77). It is the only published ECR benchmark that contains multimodal resources specific to cross-document coreference. Events here are specifically in the domain of Russia-Ukraine relations and annotated based on both saliency and the potential for conflicting perspectives.

ing web.archive.org. Of 480 ECB documents, 51% were located via Google search. We retrieved a total of 543 images; 235 of 982 documents had at least one associated image.

In addition to the overall lack of images, the retrieved document-level images may be poor representatives of individual event mentions, leading to the sparsity problem mentioned in Sec. 1. Therefore, we used Stable Diffusion (Rombach et al., 2022) to generate more relevant images and provide enough data to explore the contribution of multimodal information to ECR. Photo-realistic images were generated using sentences from ECB+ as prompts. Since a sentence can refer to multiple events, we provided an additional signal in the prompt by marking the event trigger with special tokens (<m> and </m>).

Image Encoding To encode all images as vector representations, we used three variations of Vision Transformers (ViT; Dosovitskiy et al. (2021), BEiT; Bao et al. (2021), and SWIN; Liu et al. (2021)), as well as CLIP (Radford et al., 2021). Resulting representations were the pooled output of the first-token representations from the last encoder layer for the image sequence, akin to the [CLS] token in BERT variants. Encoding the images through distinct embedding spaces decoupled them from the original language inputs.

Linear Projection Technique To project image and text representations across modalities, we first created a concatenated 3,072D (768×4) representation for an image/text pair. These concatenated representations contained the paired representation, the individual mention representations (Arg1 and Arg2), and their element-wise product (in that order). Separate concatenated representations were constructed for each modality (see Fig. 1).[3] We then used a ridge regressor to calculate the linear coefficients by minimizing the squared distances between concatenated representations from each modality for the training set. This gave us two square (3,072×3,072) "bridge" matrices: $M_{\mathcal{LLM}\rightarrow\mathcal{V}}$ and $M_{\mathcal{V}\rightarrow\mathcal{LLM}}$. We hypothesized that this bidirectional map retains crucial semantic information that a structure-preserving linear map would transfer between the two modalities. At evaluation, we matrix-multiplied the test concatenated representations with these matrices while maintaining the directionality of the linear map. These mapped representations were fed into a pairwise-scorer to get coreference clusters (see Fig. 1).

[3]All language representations came from the pretrained Longformer model (Beltagy et al., 2020).

Model Training and Fine-Tuning Following Humeau et al. (2020); Cattan et al. (2021), *i.a.*, we trained separate pairwise scorers $P_{\theta,\theta'}$: $(AB, BA)\rightarrow S_1, S_2$ on ECB+ and AIDA Phase 1. Here AB and BA are the 3,072D combined representations in A→B and B→A directions respectively, and θ and θ' are the parameters of the pairwise scorer and the LLM, respectively. This output two scores for each directional encoding, each representing the probability that the event mention pair was coreferent.[4] Thereafter, we used the CoVal Scorer (Moosavi et al., 2019) to form the final coreference clusters after applying transitive closure to identify the connected components with a threshold of 0.5 for all models. We used the same pairwise-scorer for all linear maps.

For a direct multimodal comparison, we fine-tuned fused-modality models. We concatenated the image representations with the text representations and trained four separate pairwise scorers for each combination. Due to data sparsity of real images, we only trained fused models using generated event-centric images. Training took roughly 1.0 and 1.5 hours per epoch for the LLM and the fused models, respectively. For comparison, linear mapping took ∼3s to learn a mapping between modalities. Fig. 2 shows log GPU seconds required for pairwise encoding for text, image, and fused modalities vs. bidirectional linear projection.

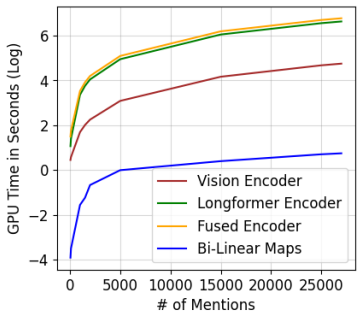


Figure 2: Pairwise encoding time in GPU seconds (log-scale on y-axis) for text (Longformer), vision (ViT), and fused models vs. Bidirectional Linear Mapping (Lin-Sem) as a function of the number of train pairs in ECB+.

3.1. Categorizing Mention Pair Difficulty

To empirically evaluate the contribution of cross-modal information toward resolving challenging event mention pairs, we used the gold-standard coreference labels to categorize unseen pairs

[4]As a human reader would likely make a consistent coreference decision regardless of which event description she read first, we used the mean of the two scores as the final probability score for training and inference.

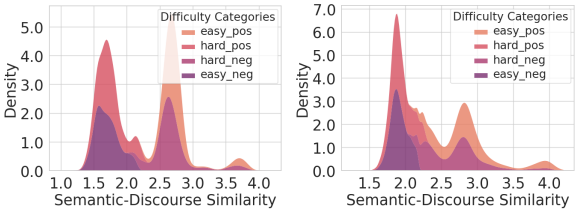


Figure 3: Kernel Density Estimation plots of semantic-discourse similarity scores (including Wu-Palmer similarity) for mention pair difficulty categories in ECB+ (L) and AIDA Phase 1 (R), showing a clear demarcation of easy and hard pairs in positive and negative labels. `easy_pos` and `hard_neg` pairs have a high semantic similarity distribution while `easy_neg` and `hard_pos` pairs have lower semantic similarity distribution.

at inference as *easy* or *hard* based on semantic and discourse-level similarities. For semantic similarities, we use Wu-Palmer Similarity (Wu and Palmer, 1994), and cosine similarity metrics. For discourse-level similarities, metadata in both datasets provides information about within-topic and within-document events which we used to score event similarities. For instance, an event pair within the same document and topic would get the highest discourse-level similarity score. These combined semantic and discourse similarity scores were then bucketed into *easy* and *hard* semantic transfer categories based on the means of coreferring and non-coreferring samples (see Fig. 3).

An example "hard" mention pair from ECB+, involving pronominal coreference, is (1) "*In a move* ***<m>*** *that* ***</m>*** *will expand its services division, Hewlett-Packard will acquire EYP Mission Critical Facilities*" and (2) "*HP to* ***<m>*** *Acquire* ***</m>*** *Data Center Consultants.*" This categorization allowed us to identify cases where multimodal features are distinctly useful based on proportion of correctly resolved *hard* pairs (see Sec. 4). Table 7 in Appendix A shows examples of *easy* and *hard* pairs for coreferring and non-coreferring samples and their respective counts.

Computation of Semantic Difficulty Categories
It is important to note that the "hard" and "easy" categories include both positive (coreferent) and negative (non-coreferent) samples. These categories are computed based on the assumption that easier coreferent (easy positive) samples should ideally have a higher overall similarity than harder ones, both in terms of semantics and at the topic and discourse level. Similarly, easier non-coreferent samples (easy negative) should ideally have a lower overall similarity. Hard coreferent (hard positive) pairs have lower overall similarity and hard non-coreferent (hard negative) pairs have higher overall similarity when compared to easy pairs of the same label.

Overall similarity for a given pair is computed as the sum of four individual scores:

1. whether a pair comes from the same topic (1 for within-topic, 0 for not),
2. whether a pair comes from the same document (1 for within-doc, 0 for not),
3. the Wu-Palmer similarity of the trigger tokens in a pair, and
4. the average cosine similarity of the vectors for the two sentences when encoded in both directions using the text-only, finetuned LLM (Longformer), inspired by (Ahmed et al., 2023).

For computing the cosine similarity scores, we take two mention-containing sentences *A* and *B* and cross-encode sentence *A* in context before sentence *B* and sentence *B* in context after sentence *A*. We then take the cosine similarity between these two encoded vectors. The positions of *A* and *B* are then reversed and they are again encoded with cross-attention in the same way. Because cross-attention is used, this results in different positional encodings for the two sentences and therefore a different cosine similarity value than the first calculation, so these values are then averaged for the final score.

Adding the aforementioned four scores gives us the final similarity scores for each pair in each label category (positive and negative). If the final similarity score for an individual positive pair is more than the mean final similarity score for all positive pairs, such a pair is categorized as easy positive. If it is less than this value, it is categorized as hard positive. On the other hand, if the final similarity score for an individual negative pair is more than the mean final similarity score for all negative pairs, the pair is categorized as hard negative, and if it is less than this value, it is categorized as easy negative.[5] The plots in Fig. 2 show the differences in the distributions of different sample categories vs. the calculated similarity scores for both the corpora. See Appendix A for more details with computed examples.

We use the gold coreference labels to obtain the label categories. However, since this categorization is only used as an evaluation tool for the initial round of experiments and then frozen for the

[5]The average final similarity for all positive samples over the ECB+ corpus is 2.25, and the average final similarity for all negative samples is 2.14. We assume AIDA Phase 1 comes from a disparate distribution, and so we categorize the difficulty of pairs in it independently using the same procedure.

ensembling experiments, the difficulty category-related information is never used during model training.

4. Results and Analysis

We evaluate using established coreference metrics (Moosavi et al., 2019), e.g., MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_e$, and CoNLL F1 (the average of MUC, B^3 and $CEAF_e$ F1) scores.

4.1. ECB+

We present results from Held et al. (2021) as a current, commonly accepted SOTA on ECB+, and from Ahmed et al. (2023), whose computationally-efficient pruning heuristic based on surface lemma similarity we follow to allow us to perform multiple experiments on a smaller compute budget. Direct comparison to text-only model (LLM) performance should be taken as a comparison to Ahmed et al. (2023) due to the preprocessing. Table 2 shows detailed results.

Models	MUC	B^3	$CEAFe$	CoNLL
Held et al. (2021)	87.5	86.6	82.9	85.7
Ahmed et al. (2023)	**90.8**	**86.7**	**84.7**	**87.4**
ViT-real→LLM	6.9	63.1	55.1	41.7
BEiT-real→LLM	87.3	80.3	76.7	81.4
SWIN-real→LLM	87.6	79.7	76.5	81.3
CLIP-real→LLM	24.7	66.3	57.5	49.5
LLM→ViT-real	88.2	80.1	77.5	81.9
LLM→BEiT-real	88.3	80.0	77.4	81.9
LLM→SWIN-real	87.9	80.3	77.8	82.0
LLM→CLIP-real	88.3	80.0	77.4	81.9
ViT-gen ⊕ LLM	85.1	86.1	80.7	84.0
BEiT-gen ⊕ LLM	82.2	84.9	78.1	81.7
SWIN-gen ⊕ LLM	82.5	85.1	78.7	82.1
CLIP-gen ⊕ LLM	89.3	84.2	82.6	85.4
ViT-gen→LLM	77.4	78.8	71.5	75.9
BEiT-gen→LLM	77.8	79.8	73.7	77.1
SWIN-gen→LLM	79.5	79.6	73.4	77.5
CLIP-gen→LLM	83.0	82.1	76.3	80.5
LLM→ViT-gen	88.1	80.0	77.2	81.8
LLM→BEiT-gen	88.3	80.0	77.4	81.9
LLM→SWIN-gen	88.2	80.1	77.4	81.9
LLM→CLIP-gen	88.3	80.0	77.4	81.9

Table 2: MM-CDCR F1 scores for MUC, B^3, $CEAFe$ and CoNLL on ECB+ test set, using LLM only, Lin-Sem ("→"), and domain-fused finetuned versions ("⊕"). Cited works are previous benchmarks on text-only CDCR. **Bold** indicates the best performer on each metric. "-real" indicates that the vision space was encoded with real images, while "-gen" indicates generated images.

Text-only vs. Multimodal Models Despite the extra training time incurred in training a fused-modality model with concatenated features (see Fig. 2), we see that the performance of the fused multimodal models does not exceed that of the text-only model (Longformer using Ahmed et al. (2023)'s preprocessing heuristic). Interestingly, the performance gap between linearly-mapped systems and fused modality models is often quite small, despite the higher compute cost of training the fused model. For instance, LLM→BEiT-gen and LLM→BEiT-real (Longformer embeddings mapped into BEiT space) slightly best the CoNLL F1 score of BEiT-gen ⊕ LLM, and BEiT-real→LLM is only 0.5 F1 points lower. Similar trends hold when comparing other fused modality models and their linearly-mapped counterparts, such as LLM→SWIN-gen, LLM→SWIN-real, and SWIN-real→LLM vs. SWIN-gen ⊕ LLM.

Semantic Transfer Categories In the coreference domain, one weakness of the CoNLL F1 metric is that specific evaluation metric-level details are obfuscated—this can be seen in Table 3: although the aforementioned examples achieve comparable CoNLL F1 scores, the linear mappings achieve a much higher MUC and B^3 recall, but lower precision, than the comparable fused models. Therefore, we do a proportional analysis of the correctly inferred (true positive) and misclassified (false positive and false negative) samples within the semantic transfer categories (see Table 4). These categorization labels were not used as supervision at any stage of training, fine-tuning, or mapping, and so an analysis of which models do better at which categories can illuminate different properties of the models, despite similar numerical performance. Table 4 shows the proportion of each result category per model, of samples that would be considered "hard" according to the mention pair difficulty categorization described in Sec. 3.

Models	MUC		B^3	
	R	P	R	P
LLM→ViT-gen	98.7	79.6	97.6	67.7
LLM→BEiT-gen	99.1	79.6	97.9	67.7
ViT-gen ⊕ LLM	80.9	89.7	85.4	86.9
BEiT-gen ⊕ LLM	75.9	89.7	82.5	87.5

Table 3: MUC and B^3 precision and recall comparison between linear mappings and comparable fused models.

Within true positives (TP), linearly-mapped models, using both real and generated images, tended to correctly retrieve a higher proportion of hard pairs compared to the text-only and fused models. For instance, for generated images, the hard sample proportion retrieved by text-to-image models is almost 4 percentage points higher than that of text-only or fused models, while image-to-text models, though lower on average, still also correctly re-

trieve a higher proportion of hard pairs. This effect appears slightly more pronounced on average in the case of real images (avg. 51.8% hard pairs in TPs, compared to 50.1% for generated images, and 46.6% for text-only).

Models	TP-Hard	FP-Hard	FN-Hard
ECB+			
Ahmed et al. (2023)	0.466	0.521	0.607
ViT-real→LLM	0.625	0.250	0.506
BEiT-real→LLM	0.521	0.436	0.434
SWIN-real→LLM	0.510	0.451	0.407
CLIP-real→LLM	0.476	0.536	0.508
LLM→ViT-real	0.507	0.456	0.441
LLM→BEiT-real	0.506	0.000	0.000
LLM→SWIN-real	0.496	0.438	0.700
LLM→CLIP-real	0.505	0.452	0.708
ViT-gen ⊕ LLM	0.432	0.591	0.635
BEiT-gen ⊕ LLM	0.437	0.606	0.584
SWIN-gen ⊕ LLM	0.404	0.620	0.642
CLIP-gen ⊕ LLM	0.477	0.506	0.729
ViT-gen→LLM	0.487	0.472	0.521
BEiT-gen→LLM	0.471	0.445	0.525
SWIN-gen→LLM	0.548	0.433	0.478
CLIP-gen→LLM	0.483	0.490	0.534
LLM→ViT-gen	0.505	0.449	0.541
LLM→BEiT-gen	0.506	0.451	0.000
LLM→SWIN-gen	0.505	0.452	0.531
LLM→CLIP-gen	0.506	0.451	0.632
AIDA Phase 1			
LLM	0.561	0.385	0.695
ViT-real→LLM	0.609	0.368	0.734
BEiT-real→LLM	0.661	0.328	0.629
SWIN-real→LLM	0.660	0.327	0.636
CLIP-real→LLM	0.627	0.332	0.657
LLM→ViT-real	0.643	0.346	0.929
LLM→BEiT-real	0.638	0.352	0.749
LLM→SWIN-real	0.667	0.333	0.562
LLM→CLIP-real	0.648	0.341	0.000

Table 4: Table showing the proportion of hard event pairs within the true positive (TP), false positive (FP) and false negative (FN) samples based on semantic transfer category (Sec. 3) for ECB+. Values of 0 indicate that no cases fit this category, resulting in zero numerator.

Ensembling Models The apparent facility of different models at correctly retrieving mention pairs of different semantic difficulties led to a question: since the mention pair difficulty was never used during training, fine-tuning, or mapping, and only as an analytic tool, could we split the mention pairs according to their difficulty, and use the different model types to handle mention pairs they on average appear to be better at? We therefore built an ensembling approach using the text-only model to handle easier pairs, and performed a grid-search through different combinations of the previously-trained multimodal models to handle harder pairs. We allowed for different multimodal models to potentially handle hard-positive pairs and hard-negative pairs and used the combined results from all models to compute the coreference metrics. Table 5 shows the best performing ensembles.

Models	MUC	B^3	$CEAFe$	CoNLL
Held et al. (2021)	87.5	86.6	82.9	85.7
Ahmed et al. (2023)	90.8	86.7	84.7	87.4
ViT-gen ⊕ LLM + LLM	89.1	86.5	84.8	86.8
BEiT-gen ⊕ LLM + LLM	87.5	85.7	83.9	85.7
SWIN-gen ⊕ LLM + LLM	87.5	85.9	83.8	85.7
CLIP-gen ⊕ LLM + LLM	90.1	85.3	83.8	86.4
ViT-gen→LLM + LLM→BEiT-gen + LLM	90.8	85.2	84.8	86.9
BEiT-gen→LLM + LLM→BEiT-gen + LLM	91.3	85.5	86.5	87.8
SWIN-gen→LLM + LLM→BEiT-gen + LLM	90.4	84.4	83.8	86.2
CLIP-gen→LLM + LLM→BEiT-gen + LLM	91.2	85.3	85.7	87.4
LLM→ViT-gen + LLM→BEiT-gen + LLM	88.7	82.3	79.4	83.5
LLM→BEiT-gen + LLM	88.7	82.2	79.1	83.3
LLM→SWIN-gen + LLM→BEiT-gen + LLM	88.7	82.2	79.1	83.3
LLM→CLIP-gen + LLM→BEiT-gen + LLM	88.7	82.2	79.1	83.3
ViT-real→LLM + LLM→BEiT-real + LLM	**94.5**	**89.5**	**91.8**	**91.9**
BEiT-real→LLM + LLM→BEiT-real + LLM	88.9	82.4	79.7	83.7
SWIN-real→LLM + LLM→BEiT-real + LLM	88.7	82.2	79.1	83.3
CLIP-real→LLM + LLM→BEiT-real + LLM	94.3	89.3	91.6	91.7
LLM→ViT-real + LLM→BEiT-real + LLM	88.7	82.3	79.3	83.4
LLM→BEiT-real + LLM	88.7	82.2	79.1	83.3
LLM→SWIN-real + LLM→BEiT-real + LLM	89.0	82.7	80.1	83.9
LLM→CLIP-real + LLM→BEiT-real + LLM	88.7	82.2	79.1	83.3

Table 5: MM-CDCR MUC, B^3, $CEAFe$ and CoNLL F1 results on ECB+ test set, using ensemble models. Format follows Table 2. Ensemble model names follow the format *Hard-N model + Hard-P model + Easy pairs model*. LLM was always used to handle Easy pairs. The best performing models for hard negative and hard positives were found using a grid search through different combinations of multimodal models. If only one model besides LLM is listed, that model was used to handle all Hard pairs.

Our best performing ensemble model used ViT-real→LLM to handle hard negative pairs, LLM→BEiT-real, to handle hard positive pairs, and the text-only language model to handle easy pairs. This resulted in a CoNLL F1 score of 91.9, with scores of 89.5 or higher across all components of MUC, B^3, or $CEAF_e$ metrics, showing the ability of this ensemble to score highly on, and balance, multiple measurements. Other ensembles, such as a variant that used CLIP-real→LLM to handle hard negatives, performed at a similar level. Two particularly interesting points emerge: 1) Using both real and generated images, LLM→BEiT routinely performed best at handling hard positive pairs; 2) Many ensemble models using Lin-Sem, especially those using a $\mathcal{V} \rightarrow \mathcal{LLM}$ mapping for hard negatives and an $\mathcal{LLM} \rightarrow \mathcal{V}$ mapping for hard positives, outperform the fused model/text-only model ensembles, despite the simplicity of the linear transformation. This suggests that not only can visual information be leveraged for correct coreference of semantically more difficult mention pairs, but also that visual information may contain fine-grained cues useful for splitting mention pairs while linguistic information is more useful to cluster them.

4.2. AIDA Phase 1

Table 6 presents a novel baseline on the multimodal AIDA Phase 1 data. This data contains unique challenges, such as a train set that is smaller than the test data, and event descriptions from sources with conflicting perspectives, explicitly addressing the ambiguity and perspective conflict challenges from Sec. 1. Since this data comes with images mappable to individual event mentions, we evaluate using only the provided images.

As with ECB+, we find that models using linear mappings compete with or slightly outperform the text only model. Using the same proportional analysis of correct and misclassified samples by difficulty category, we find that linearly-mapped models are also more likely than the text-only to resolve hard pairs correctly on this dataset (avg. hard pairs in TPs: 63.9% for $\mathcal{V} \rightarrow \mathcal{LLM}$, 64.9% for $\mathcal{LLM} \rightarrow \mathcal{V}$, and 56.1% for text-only).

We then applied the same ensembling approach to the AIDA data, using the same combination of linear mappings and the LLM according to the difficulty of the mention pair. Again we find that an ensemble model using a $\mathcal{V} \rightarrow \mathcal{LLM}$ mapping for hard negatives and an $\mathcal{LLM} \rightarrow \mathcal{V}$ mapping for hard positives performs best, although this time the model using CLIP-real→LLM as the hard negative handler comes out on top.

Models	MUC	B^3	$CEAFe$	CoNLL
LLM	80.7	49.5	54.1	61.4
ViT-real→LLM	85.9	38.4	52.7	59.0
BEiT-real→LLM	85.7	42.6	57.9	62.1
SWIN-real→LLM	82.9	46.4	55.8	61.7
CLIP-real→LLM	78.5	**52.4**	53.5	61.5
LLM→ViT-real	86.3	37.3	52.7	58.8
LLM→BEiT-real	85.7	40.2	53.1	59.7
LLM→SWIN-real	86.2	39.1	54.4	59.9
LLM→CLIP-real	86.2	37.1	52.3	58.5
ViT-real→LLM + LLM→BEiT-real + LLM	86.2	39.6	54.4	60.1
BEiT-real→LLM + LLM→BEiT-real + LLM	**87.1**	42.1	60.4	63.2
SWIN-real→LLM + LLM→BEiT-real + LLM	**87.1**	42.5	60.5	63.4
CLIP-real→LLM + LLM→BEiT-real + LLM	**87.1**	43.8	**62.8**	**64.6**
LLM→ViT-real + LLM→BEiT-real + LLM	86.2	39.0	53.5	59.6
LLM→BEiT-real + LLM	85.8	40.8	54.1	60.2
LLM→SWIN-real + LLM→BEiT-real + LLM	86.6	40.7	56.6	61.3
LLM→CLIP-real + LLM→BEiT-real + LLM	86.2	39.0	53.5	59.6

Table 6: MM-CDCR MUC, B^3, $CEAFe$ and CoNLL F1 results on AIDA Phase 1 Eval set. Format follows Tables 2 & 5. LLM denotes Longformer evaluated with Ahmed et al. (2023)'s methodology.

5. Discussion

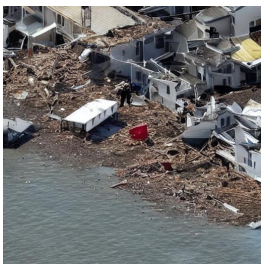
Some specific example pairs where the text-only and fused models fail to link the pair, but ensembles correctly do so, expose certain features crucial for event coreference that are present in visual information and linearly transferable, but missing in text alone or scrambled during model fusion.

ECB+ ECB+ examples of this kind include event pairs that require some sense of visual grounding, temporal logic (Schank and Abelson, 1975; Ravi et al., 2023) or pronominal context to resolve. For instance, pairs with pronominal antecedents and misleading lexical overlap like "*...dozens of others were seriously injured in the quakes, which also sent small tsunamis...*" and "*...injured in the earthquakes which rekindled bitter memories of similar deadly quakes...*"[6] were missed by the LLM and fused models. Visual cues, such as damaged buildings or injured people (either in images generated using mentions as prompts, or already present in images in news articles) can help make the link. The aforementioned example is shown in Fig. 4, and the images are generated according to the ECB+ augmentation methodology (Sec. 3). Also in Fig. 4, *Steven Moffat* and *his* appear to be ambiguously overlapping to the text-only model, which missed the event mentions that are actually about Peter Capaldi. The two facial images, which are real images associated with the event mentions, help make the link.

AIDA Phase 1 Coreferent event mentions in the AIDA dataset are notable for conflicting information, and we find cases such as "*Calling people tell about people that are jumping out of the burning building.*" vs. "*Forty-two people trapped by a fire on the third floor of the stately, Soviet-era Trades Unions building burned, suffocated or jumped to their deaths.*" Text-only fails to link ambiguous event triggers, but the images associated with each show the Trades Unions building in Odesa. In such context-sensitive pairs, the paired visual representations (image domain Arg1 and Arg2 in Fig. 1) in Lin-Sem help resolve the coreference by capturing less ambiguous information from images while the text-only pairwise scorer found low contextual similarity between the event triggers. Similarly, we see pairs with ambiguous context or pronominal anaphora, e.g., "*Buzina, 45, was shot dead*" vs. "*He was murdered*", are frequently missed by the LLM, but not by the ensemble models. In the case of this mention pair, both associated articles contain (different) pictures of the same individual, Oles Buzina, which, as with the ECB+ Peter Capaldi example, aids in the coreference[7]. Generally, for challenging corpora like AIDA Phase 1, we find visual features like faces, or background cues like angry protesters, press conferences, etc., act as cues for correctly resolving that pair.

[6] "[E]arthquakes" vs. "quakes" is misleading lexical overlap as they refer to different earthquakes. The actual event triggers are underlined.

[7] McNeely-White et al. (2022) present strong evidence for the particular effectiveness of linear transformation in face recognition.



A young girl was killed and dozens of others were seriously injured in the quakes , **<m> which </m>** also sent *small tsunamis* into Japan 's southeastern coast.

Atururi said a 10-year-old girl was killed and at least 40 people were injured in the **<m> earthquakes </m>** , which rekindled bitter memories of similar deadly *quakes* that hit the town in 2002.

Doctor Who has finally selected its 12th doctor : Peter Capaldi is officially set to replace exiting star Matt Smith as the TARDIS leader , producer *Steven Moffat* **<m> announced </m>** on the live BBC special Doctor Who Live : The Next Doctor Sunday.

Scottish actor best known for *his* role as Malcolm Tucker in The Thick of It **<m> revealed </m>** as 12th actor to play the Doctor.

Figure 4: Sample coreferent event pairs from ECB+ that were correctly linked by our best multimodal ensemble (ViT-real→LLM + LLM→BEiT-real + LLM), but not by the text-only model. Event-triggers are highlighted in yellow and text in italics illustrates lexical ambiguity or misleading lexical overlap.

6. Conclusion

In this paper we have demonstrated the utility of multimodal information in cross-document event coreference. In particular, our results demonstrate that multimodal information is useful for resolving mention pairs whose triggers have low semantic and discourse-level similarity, rendering them difficult for text-only models. We developed a method (Lin-Sem) for using linear transformations between embedding spaces to transfer semantic information between vision and language representation spaces, and used this technique in a model ensembling approach that used Lin-Sem models to handle harder mention pairs and a text-only model for easier pairs. We applied this approach to the popular ECB+ benchmark and established a novel baseline on the challenging, and explicitly multimodal, AIDA Phase 1 dataset (Tracey et al., 2022). Our best performing models beat text-only performance on these datasets by ∼3 F1 points and establish an upper bound on CDCR performance given the preprocessing used. Our ablation studies show that ensemble systems built upon our mention pair difficulty categories and using structure preserving linear maps can leverage event-specific visual cues to make correct coreference decisions about difficult mention pairs. These visual cues are of course absent in text only models, and are likely scrambled during standard multimodal fusion approaches. As such, our results present a strong case for the utility of multimodal information in NLU tasks like event coreference and argue for future increased development of such resources. Upon publication, we will release our processing pipeline and the generated/scraped images associated with ECB+.[8]

Our results should be considered in the context of our preprocessing assumptions. We use a computationally-efficient pruning heuristic that allowed us to run the high volume of experiments we showcased on a lower compute budget, while demonstrating the utility of multimodal features for coreference. Our binary semantic transfer categories (*easy*/*hard*) do not currently account for semantic similarity between pairs that cross subtopics since corpora like the ECB+ corpus do not contain coreference annotations across sub-topics (Bugert et al., 2021). However, our framework can be easily expanded to corpora like FCC (Bugert et al., 2020), with cross-subtopic events.

7. Future Work

Future directions in this line of research include exploring the feasibility of using multimodal cues to align/enhance representation spaces of monolingual LLMs, like the English-only Longformer, for Russian and Ukrainian mention pairs in the AIDA Phase 1 corpus. Given the efficiency of linear transformations and the rarity of coreference-specific parallel corpora, this may help alleviate the compute budgets needed for multilingual LLM pretraining for CDCR. Another interesting direction is evaluating our method for other challenging CDCR datasets like FCC (Bugert et al., 2020) which contains cross-subtopic events or the GVC (Vossen et al., 2018) where the SOTA is lower compared to benchmarks like ECB+. Lastly, this work represents a novel cross-modal case where affine transformations between embedding spaces has been shown to be useful (cf. McNeely-White et al. (2022); Nath et al. (2022); Merullo et al. (2023); Ghaffari and Krishnaswamy (2023)). Future work in this area entails a theoretical exploration of the properties of embedding spaces with a goal of finding performance guarantees where affine transformations successfully preserve information for different AI tasks.

[8]The AIDA Phase 1 data must be properly obtained from the Linguistic Data Consortium.

Ethics Statement

Our ablation studies required a non-trivial computation budget and concomitant resource usage, especially for the fused models with larger scoring heads on top of the LLM. Moreover, even though our `Lin-Sem` framework is substantially compute-efficient, it still required cross-modal model encoding in generating representations for deploying our linear maps between them. The images generated for this task with diffusion models might reflect social, racial, or gender-based stereotypes as are commonly seen in large generative models. Due to the nature of the AIDA Phase 1 data's focus on Ukrainian-Russian conflict, the events described therein are likely to be distressing to some.

Acknowledgements

This research was supported in part by grant award FA8750-18-2-0016 from the U.S. Defense Advanced Research Projects Agency (DARPA) to Colorado State University and the University of Colorado, and by a subcontract to the University of Colorado on grant award FA8750-19-2-1004 from DARPA. Views expressed herein do not reflect the policy or position of the Department of Defense or the U.S. Government. All errors are the responsibility of the authors.

Bibliographical References

Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin, and Nikhil Krishnaswamy. 2023. $2 * n$ is better than n^2: Decomposing event coreference resolution into two tractable problems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1569–1583, Toronto, Canada. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

Amit Bagga and Breck Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *Coreference and Its Applications*.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv e-prints*, pages arXiv–2004.

Michael Bugert, Nils Reimers, Shany Barhom, Ido Dagan, and Iryna Gurevych. 2020. Breaking the subtopic barrier in cross-document event coreference resolution. In *Text2story@ ecir*, pages 23–29.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora. *Computational Linguistics*, 47(3):575–614.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107, Online. Association for Computational Linguistics.

Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. 2021. Joint multimedia event extraction from video and article. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 74–88.

Lin Chen, Anruo Wang, and Barbara Di Eugenio. 2011. Improving pronominal and deictic coreference resolution with multi-modal features. In *Proceedings of the SIGDIAL 2011 Conference*, pages 307–311.

Zheng Chen and Heng Ji. 2009. Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 54–57, Suntec, Singapore. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552.

Agata Cybulska and Piek Vossen. 2015. Translating granularity of event slots into features for event coreference resolution. In *Proceedings*

of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Jacob Eisenstein and Randall Davis. 2006. Gesture improves coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 37–40, New York City, USA. Association for Computational Linguistics.

Andrew Finch, Young-Sook Hwang, and Eiichiro Sumita. 2005. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In *Proceedings of the third international workshop on paraphrasing (IWP2005)*.

Sadaf Ghaffari and Nikhil Krishnaswamy. 2023. Grounding and distinguishing conceptual vocabulary through similarity learning in embodied simulations. In *Proceedings of the 15th International Conference on Computational Semantics*.

Danfeng Guo, Arpit Gupta, Sanchit Agarwal, Jiun-Yu Kao, Shuyang Gao, Arijit Biswas, Chien-Wei Lin, Tagyoung Chung, and Mohit Bansal. 2022. Gravl-bert: Graphical visual-linguistic representations for multimodal coreference resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 285–297.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Huma Jamil, Yajing Liu, Nathaniel Blanchard, Michael Kirby, and Chris Peterson. 2023. Leveraging linear mapping for model-agnostic adversarial defense. *Frontiers in Computer Science*, 5.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2018. Resolving event coreference with supervised representation learning and clustering-oriented regularization. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 1–10.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623.

Karel Lenc and Andrea Vedaldi. 2015. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. 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.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

David McNeely-White, J Ross Beveridge, and Bruce A Draper. 2020. Inception and resnet

features are (almost) equivalent. *Cognitive Systems Research*, 59:312–318.

David McNeely-White, Ben Sattelberg, Nathaniel Blanchard, and Ross Beveridge. 2022. Canonical face embeddings. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space. In *International Conference on Learning Representations*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4168–4178, Florence, Italy. Association for Computational Linguistics.

Abhijnan Nath, Rahul Ghosh, and Nikhil Krishnaswamy. 2022. Phonetic, semantic, and articulatory features in Assamese-Bengali cognate detection. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 41–53, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwartz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1700–1716.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Roger C Schank and Robert P Abelson. 1975. Scripts, plans, and knowledge. In *IJCAI*, volume 75, pages 151–157.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Meihan Tong, Shuai Wang, Yixin Cao, Bin Xu, Juanzi Li, Lei Hou, and Tat-Seng Chua. 2020. Image enhanced event detection in news articles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9040–9047.

Jennifer Tracey, Ann Bies, Jeremy Getman, Kira Griffitt, and Stephanie Strassel. 2022. A study in contradiction: Data and annotation for aida focusing on informational conflict in russia-ukraine relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1831–1838.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Marc Vilain, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. 2013. Hoggles: Visualizing object detection features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't annotate, but validate: A data-to-text method for capturing event data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138.

Nishant Yadav, Nicholas Monath, Rico Angell, and Andrew McCallum. 2021. Event and entity coreference using trees to encode uncertainty in joint decisions. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 100–110.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2022. Pairwise representation learning for event coreference. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 69–78.

A. Examples of Semantic Difficulty Categories

To illustrate how the combined scores of the four individual similarity scores are used to categorize the difficulty of mention pairs, consider an ECB+ cross-document event mention pair from each of the four semantic difficulty categories (Sec. 3.1).

Easy-positive (Easy-P)

- Sentence 1: Advanced Micro Devices to <m> *acquire*</m> microserver vendor SeaMicro for $334 million.
- Sentence 2: AMD to <m> *Acquire* </m> Server Start - Up

In the above event mention pair, the final similarity score is $2.607 = 1 + 0 + 1 + 0.607$. Since this is more than the average of the final similarity for the positive samples (2.25), this mention pair is categorized as easy-positive.

Hard-positive (Hard-P)

- Sentence 1: <m> *4.6 quake* </m> rattles Sonoma County early Thursday.
- Sentence 2: <m> *4-Plus Earthquake* </m> Recorded Near Healdsburg

In the above event mention pair, the final similarity score is $2.12 = 1 + 0 + 0.5 + 0.62$. Since this is less than the average of the final similarity for the positive samples (2.25), this mention pair is categorized as hard-positive.

Easy-negative (Easy-N)

- Sentence 1: Apple <m> *Unveils* </m> New Flagship Macbook Pro.
- Sentence 2: Next, global marketing VP Phil Schiller <m> *announced* </m> updates to the MacBook line.

In the above event mention pair, the final similarity score is $1.92 = 1 + 0 + 0.22 + 0.70$. Since this is less than the average of the final similarity for the negative samples (2.14), this mention pair is categorized as easy-negative.

Hard-negative (Hard-N)

- Sentence 1: <m> *4.6 quake* </m> rattles Sonoma County early Thursday.
- Sentence 2: <m> *4-Plus Earthquake* <m> Recorded Near Healdsburg

In the above event mention pair, the final similarity score is $2.60 = 1 + 0 + 1 + 0.60$. Since this is more than the average of the final similarity for the positive samples (2.14), this mention pair is categorized as hard-negative. Table 7 shows a few more examples of each type of semantic category and the total number of pairs in the ECB+ test set that fall into that category.

B. Further Details on Definitions and Semantic Equivalence

Our definition of semantic equivalence is more generalized than those evaluated in works like Finch et al. (2005) where semantic equivalence is specific to tasks like machine translation.

Although the approximate invertibilty of image representation functions like HOG (Vondrick et al., 2013) has been proven, our assumption of bijectivity of both $\mathcal{V}$ and $\mathcal{LLM}$ (see Sec. 3, **Semantic Equivalence**) is based on the fact that Transformer-based bidirectional encoders like the Longformer (Beltagy et al., 2020) are still arguably more bijective than static embeddings like word2vec (Mikolov et al., 2013) since the latter may have a many-to-one correspondence.

C. Further Details on Image Generation and Vision Encoding

To minimize GPU-compute requirements for image generation, we only generate one high-quality image for each event-mention in the ECB+ corpus. Our linear projection technique lets us encode events individually as well as pairwise without expending additional resources for generating image representations of paired mentions.

For generating high-quality, photo-realistic images we chose a guidance scale of 7.5. This enabled us to generate images that were both creative and relevant to coreference-specific natural language descriptions. In order to ensure a more balanced trade-off between efficiency and quality, we set the number of inference steps to 15. The resulting RBG images had a resolution of 512×512. The image generation process for the entire ECB+ corpus (6,833 event mentions) took ∼4 hours on an NVIDIA GeForce RTX 3090.

To obtain image embeddings, we used the four models mentioned (ViT, BEiT, SWIN, and CLIP) from HuggingFace.[9] We use the generated images to get both individual as well as paired-image embeddings from each of the Transformer-based models (excluding CLIP) after converting the original image to 224×224 resolution using the AutoImageProcessor on HuggingFace. CLIP requires an

[9] https://huggingface.co

Semantic Category	Pair1	Pair2	Pairs
Easy-P	An Oklahoma man has pleaded not guilty to two first - degree murder charges for the <m> **deaths** </m> of an Arkansas woman and her fetus.'	Oklahoma man pleads not guilty in <m> **deaths** </m> of Arkansas woman and her fetus.	2661
Hard-P	In a move <m> **that** </m> will expand its services division , Hewlett - Packard will acquire EYP Mission Critical Facilities , a New York company that offers data center consulting services.	HP to <m> **Acquire** </m> Data Center Consultants	2730
Easy-N	The UN has disputed claims that Hamas militants <m> **fired** </m> mortars from the Gaza school that has suffered the deadliest attack of the war with Israel.	Pressure to obtain a <m> **ceasefire** </m> in Gaza has been mounting , with the EU warning Israel it was " destroying " its image , while Israeli forces on Tuesday (6 January) killed at least 40 people during an attack on a United Nations-run school in Gaza.	1315
Hard-N	Atururi said a 10-year-old girl was killed and at least 40 people were injured in the earthquakes , which rekindled bitter memories of similar deadly <m> **quakes** </m> that hit the town in 2002.	As aid started to arrive , hundreds of aftershocks continued to rattle the coastal city which was hit by the 7.6 and 7.5 magnitude <m> **quakes** </m> early on Sunday , cutting power and prompting a brief tsunami warning.	1082

Table 7: Examples of easy and hard samples from our semantic transfer categories for the ECB+ test set. Suffixes P and N denote coreferring and non-coreferring pairs respectively according to the gold standard. The "Pairs" column shows the number of samples in that category.

additional linguistic supervision component along with image inputs. As such, we used sentences containing the respective event mentions while ensuring a maximum token length of 77 around the event mentions to retain context. More specifically, these embeddings are the pooled output of the classification or first-token representations from the last encoder layer for the image sequence.

D. Pairwise Scorer Hyperparameters

Our supervised pairwise scorer trained on top of the LLM is a two-layer (768 and 128 neurons) neural network trained along with the base Longformer (Beltagy et al., 2020), with binary cross entropy (BCE) loss and final sigmoid activation for 10 epochs over the annotated (gold) labels. We add two special tokens (<m> and </m>) to the LLM vocabulary while training the scorer, that trigger a specific event mention while encoding the text. This helps us generate a contextualized representation of the text akin to the [SEP] token in BERT (Devlin et al., 2018). We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of $1e-4$ for the scorer and $1e-5$ for the LLM.

We use the same hyperparameters mentioned above for our fused models, except for the input layer size (6,144 = 3,072$\times$2) of the pairwise scorers. For training both the LLM and the fused models, we utilize a single NVIDIA A100 GPU with 80GB memory.

E. Impact of Coreference Evaluation Metrics

For coreference tasks, the metric chosen has a substantial impact on the numerical results. MUC is a link-based metric that calculates the minimum number of *missing* links between mentions in the predicted entities in comparison to the gold-standard entities to get its recall and precision. It is also the least discriminative since it does not differentiate whether an extra link merges two singletons. B^3 is a mention-based metric that calculates overall recall and precision based on a combination of the recall and precision of the individual mentions. Unlike MUC, the presence of singletons in the corpus disproportionately affects B^3 scores. $CEAF_e$ intends to overcome B^3's tendency to use a mention more than once when comparing entities containing that mention. It uses an optimized mapping to align the entities in key and response to calculate its precision and recall. This may explain why the precision and recall trends in our results are inverted when using B^3/MUC vs. when using $CEAF_e$.

Table 8 and Table 9 shows detailed results including the precision and recall scores from various metrics for the ECB+ and AIDA Phase 1 corpus respectively. Empirically, the nature and the extent of such semantic transfer can also be analyzed via the MUC metric, in which text to image models performed closest to the text-only LLM

Models	MUC			B^3			$CEAFe$			CoNLL
	R	P	F_1	R	P	F_1	R	P	F_1	F_1
ViT-gen ⊕ LLM + LLM	89.2	89.0	89.1	90.5	82.9	86.5	84.7	84.9	84.8	86.8
BEiT-gen ⊕ LLM + LLM	86.5	88.6	87.5	88.5	83.0	85.7	85.2	82.7	83.9	85.7
SWIN-gen ⊕ LLM + LLM	85.7	89.3	87.5	88.0	83.8	85.9	85.8	81.8	83.8	85.7
CLIP-gen ⊕ LLM + LLM	94.8	85.9	90.1	95.0	77.3	85.3	78.5	89.8	83.8	86.4
ViT-gen→LLM + LLM→BEiT-gen + LLM	96.4	85.8	90.8	96.0	76.7	85.2	78.4	92.2	84.8	86.9
BEiT-gen→LLM + LLM→BEiT-gen + LLM	96.0	87.1	91.3	95.6	77.4	85.5	81.2	92.7	86.5	87.8
SWIN-gen→LLM + LLM→BEiT-gen + LLM	96.6	84.9	90.4	96.1	75.2	84.4	76.8	92.2	83.8	86.2
CLIP-gen→LLM + LLM→BEiT-gen + LLM	96.7	86.4	91.2	96.1	76.7	85.3	79.4	92.9	85.7	87.4
LLM→ViT-gen + LLM→BEiT-gen + LLM	96.9	81.7	88.7	96.3	71.8	82.3	70.4	90.9	79.4	83.5
LLM→BEiT-gen + LLM	**97.0**	81.6	88.7	**96.4**	71.6	82.2	70.1	90.8	79.1	83.3
LLM→SWIN-gen + LLM→BEiT-gen + LLM	**97.0**	81.6	88.7	**96.4**	71.6	82.2	70.1	90.8	79.1	83.3
LLM→CLIP-gen + LLM→BEiT-gen + LLM	**97.0**	81.6	88.7	**96.4**	71.6	82.2	70.1	90.8	79.1	83.3
ViT-real→LLM + LLM→BEiT-real + LLM	95.9	**93.2**	**94.5**	95.6	**84.1**	**89.5**	**90.2**	93.4	**91.8**	**91.9**
BEiT-real→LLM + LLM→BEiT-real + LLM	96.9	82.0	88.9	96.3	72.0	82.4	71.0	91.0	79.7	83.7
SWIN-real→LLM + LLM→BEiT-real + LLM	**97.0**	81.6	88.7	**96.4**	71.6	82.2	70.1	90.8	79.1	83.3
CLIP-real→LLM + LLM→BEiT-real + LLM	95.9	92.9	94.3	95.6	83.8	89.3	89.8	**93.5**	91.6	91.7
LLM→ViT-real + LLM→BEiT-real + LLM	**97.0**	81.8	88.7	**96.4**	71.7	82.3	70.3	90.9	79.3	83.4
LLM→BEiT-real + LLM	**97.0**	81.6	88.7	**96.4**	71.6	82.2	70.1	90.8	79.1	83.3
LLM→SWIN-real + LLM→BEiT-real + LLM	96.9	82.2	89.0	96.3	72.4	82.7	71.4	91.1	80.1	83.9
LLM→CLIP-real + LLM→BEiT-real + LLM	**97.0**	81.6	88.7	**96.4**	71.6	82.2	70.1	90.8	79.1	83.3

Table 8: MM-CDCR MUC, B^3, $CEAFe$ and CoNLL F1 results on ECB+ test set, using ensemble models. Format follows Table 2. Ensemble model names follow the format `Hard-N model + Hard-P model + Easy pairs model`. `LLM` was always used to handle Easy pairs. The best performing models for hard negative and hard positives were found using a grid search through different combinations of multimodal models. Where only one model is besides `LLM` is listed, that model was used to handle all Hard pairs.

Models	MUC			B^3			$CEAFe$			CoNLL
	R	P	F_1	R	P	F_1	R	P	F_1	F_1
LLM	83.3	78.2	80.7	81.6	35.6	49.5	50.3	58.6	54.1	61.4
ViT-real→LLM	96.4	77.5	85.9	96.3	24.0	38.4	39.1	81.1	52.7	59.0
BEiT-real→LLM	93.8	78.9	85.7	93.7	27.5	42.6	46.4	77.3	57.9	62.1
SWIN-real→LLM	87.7	78.5	82.9	86.5	31.7	46.4	48.9	65.0	55.8	61.7
CLIP-real→LLM	78.5	78.6	78.5	78.3	**39.3**	**52.4**	**53.6**	53.4	53.5	61.5
LLM→ViT-real	**97.4**	77.5	86.3	97.2	23.1	37.3	38.3	**84.7**	52.7	58.8
LLM→BEiT-real	95.7	77.6	85.7	95.3	25.5	40.2	39.9	79.2	53.1	59.7
LLM→SWIN-real	96.2	78.0	86.2	95.9	24.5	39.1	40.9	81.2	54.4	59.9
LLM→CLIP-real	**97.4**	77.4	86.2	**97.2**	22.9	37.1	37.9	84.6	52.3	58.5
ViT-real→LLM + LLM→BEiT-real + LLM	96.5	77.9	86.2	96.4	24.9	39.6	40.5	82.9	54.4	60.1
BEiT-real→LLM + LLM→BEiT-real + LLM	96.2	79.6	**87.1**	96.0	27.0	42.1	47.0	84.5	60.4	63.2
SWIN-real→LLM + LLM→BEiT-real + LLM	96.0	79.7	**87.1**	95.9	27.3	42.5	47.3	84.0	60.5	63.4
CLIP-real→LLM + LLM→BEiT-real + LLM	95.3	**80.3**	**87.1**	95.4	28.4	43.8	50.2	83.7	**62.8**	**64.6**
LLM→ViT-real + LLM→BEiT-real + LLM	96.9	77.7	86.2	96.7	24.4	39.0	39.4	83.1	53.5	59.6
LLM→BEiT-real + LLM	95.5	77.9	85.8	95.2	26.0	40.8	41.1	79.2	54.1	60.2
LLM→SWIN-real + LLM→BEiT-real + LLM	96.4	78.6	86.6	96.1	25.8	40.7	43.0	82.9	56.6	61.3
LLM→CLIP-real + LLM→BEiT-real + LLM	96.9	77.7	86.2	96.7	24.4	39.0	39.4	83.1	53.5	59.6

Table 9: MM-CDCR MUC, B^3, $CEAFe$ and CoNLL F1 results on AIDA Phase 1 Eval set. Format follows Tables 2 & 8. `LLM` denotes Longformer evaluated with Ahmed et al. (2023)'s methodology on novel data. Ensemble model names follow the format `Hard-N model + Hard-P model + Easy pairs model`. `LLM` was always used to handle Easy pairs. The best performing models for hard negative and hard positives were found using a grid search through different combinations of multimodal models. Where only one model is besides `LLM` is listed, that model was used to handle all Hard pairs.

baseline Ahmed et al. (2023) while outperforming fused models for the ECB+ corpus. Here, linear semantic transfer reduces the missing links generated after partitioning for calculating recall. Because the MUC is link-based, it calculates missing links needed to replicate the gold cluster chains. As such, fine-grained semantic transfer using linear maps leads to a high recall system with equivalent F1 scores compared to the text-only LLM even when precision takes a hit. This trend is also seen in B^3. The remaining metric, $CEAF_e$, shows an inverse trend where recall for text to image models is lower than the text-only LLM as well as fused models. However, lower recall is balanced by a increase in precision for text to image models relative to the others. In general, some semantic transfer is observed in all the three metrics albeit with some variations. An in-depth study of how the extent of semantic information transfer varies between metrics is a part of future work.

F. On CLIP as Image-Generator

We include CLIP (Radford et al., 2021) as one of the image encoders to explore the extent of linear semantic information transfer when a vision encoder is trained with linguistic supervision (unlike the other three vision encoders chosen). This

allows us to conduct more exhaustive ablation experiments to study the fidelity of such multimodal semantic transfer.

Although the latent diffusion model used for image generation also leverages a CLIP-guided image synthesis, there is no apparent information overlap between the latent spaces of the former with the CLIP encoder due to the nature of the event coreference task as well as the separate training objectives between the two. For instance, an event-coreference sample sentence may contain more than one event while captions usually contain one major event (or actions). Moreover, captions used to train CLIP are on average more precise and shorter sequences compared to event-coreference documents which include language not directly relevant for inferring that specific coreference. As such, our design choice to include CLIP as one of the vision-encoders still satisfies the requirement for the generalizability of such models for linear semantic transfer without giving it an undue advantage through overlap of the CLIP text encoder and the image generation model.