WikiScenes with Descriptions: Aligning Paragraphs and Sentences with Images in Wikipedia Articles

Özge Alaçam, Ronja Utescher, Hannes Grönner, Judith Sieker and Sina Zarrieß
Computational Linguistics, Dept. of Linguistics
University of Bielefeld, Germany
{oezge.alacam,hannes.groener,j.sieker,sina.zarriess}@uni-bielefeld.de

Abstract

Research in Language & Vision rarely uses naturally occurring multimodal documents as Wikipedia articles, since they feature complex image-text relations and implicit image-text alignments. In this paper, we provide one of the first datasets that provides ground-truth annotations of image-text alignments in multi-paragraph multi-image articles. The dataset can be used to study phenomena of visual language grounding in longer documents and assess retrieval capabilities of language models trained on, e.g., captioning data. Our analyses show that there are systematic linguistic differences between the image captions and descriptive sentences from the article's text and that intra-document retrieval is a challenging task for state-of-the-art models in L&V (CLIP, VILT, MCSE).

1 Introduction

Research in Language & Vision (L&V) aims at building models that ground language in the visual modality and therefore requires datasets that align text and images. To date, most work in L&V uses datasets that have been obtained via annotation of images in a way that image and text are aligned by construction as in, e.g., image captioning or VQA datasets (Thomee et al., 2016; Lin et al., 2014b; Young et al., 2014a). Multimodal image-text data that occurs "in the wild", as in, e.g., articles, recipes, comics, etc., is less commonly used since their image-text relations are much more complex (Bateman, 2008) and the alignment of images and text is often left implicit. Existing work on processing image-text alignment in multi-modal documents has usually been unsupervised, facing the challenge of missing evaluation and training data (Hessel et al., 2019). For this reason, it is unclear to what extent state-of-the-art (multi-modal) language models can discover text-image alignments in complex multi-image multi-paragraph documents and to what extent grounding capabilities in these models are biased by specific linguistic properties of annotated captions. With this work, we contribute to closing this gap and provide one of the first datasets that provide ground-truth annotations of image-text alignment in complex multimodal documents.[1]

Figure 1 shows a paragraph from the Wikipedia article on the *Reims Cathedral*[2], illustrating some of the complexities that can arise in text-image alignment in real multimodal documents. The paragraph contains highly descriptive sentences that refer to visual elements of the building shown in corresponding images. Thus, in this example, three sentences from the same paragraph *match* three different images, but there is no explicit alignment between sentences and images (e.g. through references). The paragraph also contains sentences that are not descriptive and do not match any of the images. At the same time, the images are accompanied by captions that briefly describe the image content and make it easier for the reader to establish its relation to the main text. Furthermore, this paragraph is embedded in a much longer document which contains many more, possibly matching images of this building. These alignment patterns between images and sentences in a longer text as well as captions of these images and corresponding sentences have, to date, not been extensively studied in L&V research and there is currently no available dataset that provides annotations for text-image alignments in Wikipedia articles.

In this paper, we conduct an annotation study on an existing dataset of multimodal Wikipedia articles on buildings, WikiScenes (Wu et al., 2021), and enrich the dataset with annotations of alignments between textual elements (sentences, paragraphs) and images. Since the articles in

[1]The dataset is available here: https://github.com/clause-bielefeld/wikiscenes_descriptions

[2]https://en.wikipedia.org/wiki/Reims_Cathedral

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 93 - 105

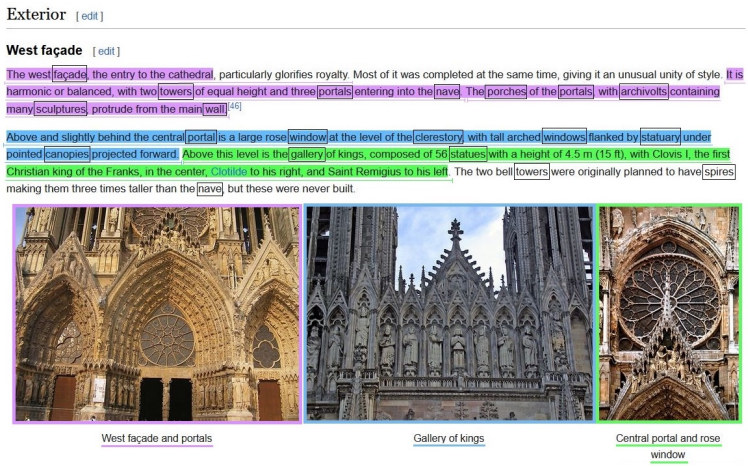


Figure 1: A highly descriptive paragraph and corresponding images from the Wikipedia article on the *Reims Cathedral*. Sentences that match an image are highlighted in the same color as the caption of the respective image.

Wikiscenes are about visual entities from the domain of historical buildings, they feature text that is at times highly descriptive and, thereby, comparable to caption-like descriptions (see, e.g., the mention of the facade of the *Reims Cathedral* in Figure 1). We restrict our annotation study to descriptive relations between text and images, i.e. textual elements that describe visual content shown in an image within the article, refraining from including more complex discourse relations involving complementary relations and others (Kruk et al., 2019). To deal with the fact that the articles are rather long and contain many images, we introduce a two-step annotation procedure, where we first ask annotators to skim the article for relations between paragraphs and images, and then annotate sentence-image alignments in a second step.

The dataset we obtain from our annotation setup, *WikiScenes with Descriptions*, can enhance research on visual language grounding in longer documents and assess grounding capabilities in language models. Our initial analyses in this paper focus on understanding how the descriptive sentences that occur within the main text and that match a particular image differ from captions of that image. We also experiment with baseline intra-document retrieval to evaluate L&V models on image-text alignment in our dataset. These analyses address the following research questions:

- Do descriptions of images in articles show different linguistic properties than captions of the corresponding images?
- Do the original captions in Wikipedia differ systematically from captions generated by captioning models?
- Can similarity-based retrieval based on the images' captions serve as a robust baseline for image-text alignment?
- How does image-sentence retrieval baselines with pretrained VILT (Kim et al., 2021) and CLIP (Radford et al., 2021) compare to caption-sentence retrieval?

Our analyses reveal systematic linguistic differences between the image captions on the one and descriptive sentences from the article's text at both linguistic and conceptual levels. We show that our dataset can serve as a challenging benchmark for image-text alignment in long documents.

2 Background

Our data collection is related to other efforts focused on multi-modal articles, e.g., WikiCaps (Schamoni et al., 2018) and WIT (Srinivasan et al., 2021), or datasets for news image captioning (Liu et al., 2020; Biten et al., 2019; Hollink et al., 2016). In comparison to these, our extension of Wu et al. (2021)'s*WikiScenes* features more detailed annotations of grounded text spans within sentences of the main text. Annotation of relations between spans or entities in longer text is generally challenging, as discussed in, e.g., work on coreference (Ghaddar and Langlais, 2016; Bamman et al., 2019). Annotation of multi-modal documents further comes with the significant complication that the number of possible combinations of text spans and images increases quadratically with the length of the text and the number of images.

There is some work on L&V datasets and tasks that capture more varied semantic or discursive relations between image and text: Kruk et al. (2019)

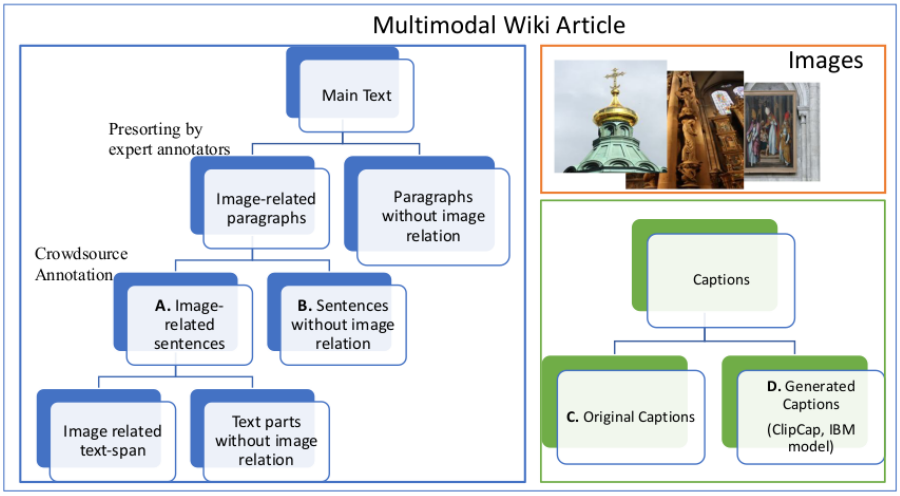


Figure 2: Illustration of the overall annotation procedure for the *WikiScenes with Descriptions* datasets, showing different levels and modalities of the annotation scheme

tag the image intent in multi-modal Twitter posts, distinguishing between intents like 'provocative', 'expressive' or 'promotive'. Their annotations assign a global label to the image which captures the relation to the text as a whole. This goes beyond literal image descriptions but still does not capture structurally diverse referential relations. Alikhani et al. (2019) investigate text-image coherence in recipe texts that describe sequences of consecutive actions in a cooking context. Structurally, the recipe text is already segmented, with an image aligned to each step. Alikhani et al. (2019) distinguish image-text relations concerning which part of the action is shown and whether all entities affected by an action are visible / mentioned in the text. Both papers work on naturally occurring texts, though these are still relatively short (tweets and 1-2 sentences per step respectively). Cheema et al. (2023) propose to combine frameworks from the area of semiotics with computational analysis of image-text relations, suggesting a framework for multimodal news analysis. In contrast to these accounts, our dataset features more or less uniform relations between texts centered on buildings and images, i.e. the texts stand in a descriptive relation to the content of images.

Muraoka et al. (2020) work with a more coarse-grained and somewhat simplified version of the problem discussed in this paper. Their task is to correctly predict the physical alignment of images and sections in Wikipedia articles. This approach utilizes the inherent document structure and consequently saves on expensive manual annotation. However, our observations call into question the presupposition that alignment in layout entails alignment in content. A similar text-image matching task is discussed in Hessel et al. (2019), where the authors seek to match the images in a document to the most relevant sentences in it (leaving out the captions). Their model is trained on collections of sentences and images from the same documents or different documents, for instances of non-relatedness. This information is used at test time to estimate the individual links between the sentences and images of a given document. Hessel et al. (2019) is highly relevant to the concerns discussed in this paper because it shows some success in handling comparatively large amounts of text in the genre of Wikipedia articles. Very recently, (Liu et al., 2023) presented the DocumentCLIP model designed to capture the interaction of text and images in longer multimodal documents. Importantly, they assume that images are, by default, aligned to the paragraph they co-occur with in the spatial document layout. This is a strong assumption and our dataset of ground-truth alignments between sentences, paragraphs, and images can be used to further test and benchmark such models.

3 Data collection

In this Section, we introduce our data collection and annotation procedures. Figure 2 shows an overview of the procedures, consisting of several stages with annotations completed at different levels, employing expert annotators and crowdsourcing. In the following, we detail each annotation stage.

3.1 Text and Paragraph Selection

From the *WikiScenes* corpus (Wu et al., 2021), we randomly sample 47 articles from the set of 98 articles. The first annotation step is a preselection of paragraphs and images that are candidates for text-image alignment. The three annotators annotated 1101 images and 1900 paragraphs. Due to the excessive number of possible paragraph-image combinations, thirty short to medium-length and one long articles were exhaustively annotated. Annotators were instructed (i) to make a snap judgment on whether a paragraph contained at least one reference to the image, (ii) to ignore non-photograph images such as plans, schemes, and paintings as well as aerial images and (iii) to consider only what is visible an image. The second and third instructions intend to exclude more complex image-paragraph correspondences and relations, that go beyond merely descriptive relations. As an example, given an image of a tower, annotators were instructed to consider sentences like *The tower was built in 1700.* as (potentially) related, while *The original altar was destroyed in the French Revolution.* is not related (even though it could be the case that the altar is inside the tower).

3.2 Fine-grained Image-Paragraph Annotations

The second annotation phase involves sentence & word-level annotations on the pre-selected paragraphs. 623 image-paragraph combinations were randomly sampled from the items collected in the previous annotation stage and evaluated by three annotators using crowd-sourcing. We recruited a group of 255 workers through Amazon Mechanical Turk. The annotators were given image and paragraph pairs, and instructed to highlight only text spans that describe something visible in the accompanying image. This ensures that the annotated text spans contain descriptions of the image or something in it. The annotation instruction are given in the Appendix, Figure 6. The average time per task was 137.6 seconds, workers were paid 0.35 $ per task.

The result of the annotation process is a collection of pairs of text spans (at sentence- and word-level) and captioned images that depict real-world objects.

Interrater agreement. At the sentence, level, if the majority of the annotators (two out of three) annotated at least one word in a sentence, the sentence is considered as depicted/matched to the respective image. We removed the cases where an annotator selects the entire paragraph instead of highlighting relevant parts. On average, the three crowd-workers who annotated each item agreed on the match or non-match of 65 % of sentences. While Wikipedia articles are aimed at a general audience, the annotation task is nonetheless non-trivial due to the complexity of the subject matter that requires a specialized vocabulary of the domain. For this reason, we believe this agreement to be of sufficient quality for further analysis. The dataset with the annotations and the generated captions at both sentence and text-span levels will be publicly available. For the rest of the paper, we present text-to-caption/image or caption/image-to-text at sentence-level alignment.

3.3 Captions

As illustrated in Figure 2, in addition to the original captions provided with the image in the wiki articles, we generated captions for the images using existing image captioning models, namely ClipCap (Mokady et al., 2021) and IBM-MAX.

ClipCap[3] (Mokady et al., 2021) is a lightweight caption generation model, based on CLIP encodings (Radford et al., 2021). It benefits from CLIP's rich semantic latent space shared by both visual and textual data trained on more than 400 M text-image pairs. In addition to the base model, we also further finetune it with several settings, the details of the finetuning are given in Appendix A.4. ClipCap-based models are listed as:

1. *clip-base*: It is the base ClipCap model without finetuning (using the CLIP Model ViT-B/32 and greedy search decoding)
2. *clip-ft*: It is created by finetuning the CLIP Image Encoder instead of the ClipCap model. 1270 unseen image-caption pairs are used for finetuning.
3. *clip-ft-gpt-20e*: It is obtained by finetuning the ClipCap model (both the prefix encoder and GPT-2 [4]

On the other hand, the IBM-MAX, inspired by Vinyals et al., 2017, does not use a transformer architecture or a large pretrained language model; instead, it utilizes an image encoder based on a

[3] https://github.com/rmokady/CLIP_prefix_caption

[4] with 10 epochs, prefix length 10, MLP Mapping with prefix size 512, lr 2e-5, with longer epochs (n=20)

deep convolutional net trained on MSCOCO images (Lin et al., 2014a), and an LSTM-based text decoder to generate the description. Both models generate a sentence describing the image content.

3.4 Data overview.

The dataset contains unique 3923 sentence-image-caption triples, with 1989 unique sentences. After the agreement analysis, we ended up with 683 matched sentences – image/caption pairs (A in Figure 2) and 1306 unmatched sentences (i.e. sentences from the same set of articles with no relation to any image (B in Figure 2).

4 Methods

This Section introduces the methods we use to analyze our dataset and to test L&V models on it. In our experiments, we look at two ways of aligning text and images: first, we study sentence-caption alignment, i.e. we investigate whether captions of images in an article are similar to sentences in the article's text that annotators marked as matching this image. Second, we study sentence-image alignment using multimodal L&V models.

4.1 Sentence – Caption Alignments

To explore the relations between sentences and captions, we investigate whether semantic similarities between image captions and matched/unmatched sentences constitute a promising baseline for automatic image-text alignments. We employ two types of sentence embeddings. First, we use text-only sentence representations extracted from the sentence transformer model (SBERT) from the Huggingface platform (Reimers and Gurevych, 2019). As the second method, we utilize pre-trained multimodal sentence representations (MCSE) provided by Zhang et al. (2022). MCSE are visually grounded sentence embeddings obtained by fine-tuning pre-trained models (e.g., ROBERTA-base (Liu et al., 2019)) in a contrastive learning framework. The sentence embeddings are enriched by training on a subset of Flickr30k (Young et al., 2014b) or MS-COCO (Lin et al., 2014b) image-caption dataset (30K images with multiple captions) and Wiki-1M text-only corpus. We used the pretrained weights using *flickr-mcse-roberta-base-uncased*[5]. We give each textual element as input to each pre-trained model and extract their CLS token embeddings.

We compute text-image alignments in two directions and with different candidate sets: we retrieve captions (or images) based on the sentence (sentence-to-caption) or retrieve the sentence given the caption (caption-to-sentence). In both cases, we distinguish between the **match** condition, where the set of candidate sentences is restricted to sentences that match at least one of the images in the article, and the **all** condition where we include all sentences, i.e. un-matched sentences that are not grounded in any of the images.

Sentence-to-caption. For this condition, the retrieval analysis is conducted by calculating the ranking of each sentence in (i) paragraph-related captions, (ii) article-related captions, and (iii) all captions in the dataset. These are referred to as *caption-sets* for the following analysis. We have also calculated the paired sentence-caption similarities and presented them in the Appendix A.5.

Caption-to-sentence. In this condition, we measure the ranking of each caption in three respective sentence sets: (i) the sentences in the same paragraph, (ii) the sentences in the same article, and (iii) all sentences in the dataset.

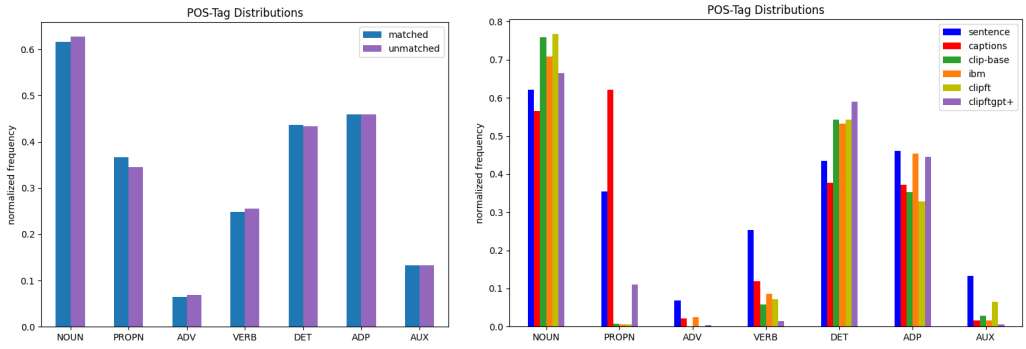
4.2 Sentence – Image Alignments

In addition to comparing the sentence embeddings among various textual elements of the articles, we also analyze the similarities between image and textual element pairs (A to D separately, see Figure 2). To obtain image–text embeddings, we employ two state-of-the-art multimodal models with zero-shot capabilities: CLIP and VILT[6] .

VILT. VILT (Kim et al., 2021) is proposed as an efficient solution for real-time image retrieval or visual question-answering tasks. It handles the modalities in a single unified manner, instead of a simple fusion of the modalities, the training algorithm utilizes a more elaborate inter-modal interaction scheme, which in return could be very valuable for more complex vision-language tasks like our case. The efficiency comes from how they process and represent the images with convolution-free encoding. It is trained in a wide variety of datasets, including MSCOCO (Lin et al., 2014b) and Flickr30K (Young et al., 2014a).

[5] https://github.com/uds-lsv/MCSE

[6] We also experimented with BLIP-2 model from the huggingface library https://huggingface.co/docs/transformers/main/en/model_doc/blip-2. Since the initial exploration indicates a similar performance to the CLIP with a longer calculation time, we abandoned it.



(a) Matched vs. unmatched sentences

(b) Sentences versus all caption types

Figure 3: POS tag distributions of matched vs. unmatched sentences, and sentences vs. captions (original Wiki captions and generated captions)

CLIP. The CLIP (Radford et al., 2021) model uses two separate encoders to embed text and images. It is trained on 400 M image–text pairs using contrastive learning utilizing Visual Transformers (Dosovitskiy et al., 2020). It is widely used for many L&V tasks, including zero-shot classification and retrieval.

Similar to the analysis of sentence-caption relations, we explored the sentence-image relations in two directions and distinguished the **match** condition (candidates restricted to matched sentences) and the **all** condition (all sentences).

Sentence-to-image. The ranking of each matched and unmatched sentence in two different sets of candidate lists to all images (i) from the same paragraph and (ii) from the same article. Due to the computational costs, we exclude the retrieval from the entire dataset for the multimodal models.

Image-to-sentence The ranking of each image in two different sets of candidate lists to all sentences (i) from the same paragraph and (ii) from the same article.

5 Results and Analysis

In this Section, we analyze the relationship between images, captions, and sentences from a linguistic and application perspective. Section 5.1 compares linguistic properties between captions and descriptions. Then we conduct experiments on intra-document retrieval using the methods for sentence–caption and sentence–image alignment in Section 4. , comparing the performance of unimodal and multimodal embedding models.

5.1 Analysis: Linguistic Differences between Sentences and Captions

To compare language use in descriptive sentences in the main text of an article to captions below images, we look at the distribution of tokens, PoS, and NER tags in sentences and captions.

Table 1 lists the number of unique captions and the average token length for each method. ClipCap produced 157 unique captions (such as *'English baroque structure on a sunny day'* for the image in Figure 1 but also the number of hallucinations or meaningless captions like *'a city in the smoke'* and *'a city is a city'* were not negligible. On the other hand, IBM-MAX generated 109 unique captions, significantly fewer compared to ClipCap. Yet, these are often visual descriptions such as *'a large building with a clock tower on top'* and *'a large cathedral with a clock on the wall'*.

As expected, the wiki captions are significantly shorter (7.43) than the sentences in the main text (28.47). ClipCap and IBM MAX models produce captions of lengths similar to the wiki captions (6.81 and 10.04). CLIP-base captions tend to be shorter, while IBM captions are slightly longer than the original captions. With CLIP fine-tuning, the generated captions get longer (8.09), but incorporating GPT-2 prefixes causes the model to generate fewer unique sentences (128). Because the main text sentences are significantly longer than any caption, the rest of the analysis is conducted on the

Table 1: Basic statistics on original and generated captions in *WikiScenes with Descriptions*

	Wiki	Clip-base	IBM	Clipft	Clip-ft-gpt-20
Unique captions	325	157	109	240	128
Average token count	7.43	6.81	10.04	7.22	8.09

normalized counts by the sentence length.

Figure 3 shows the distribution of POS and NER tags, obtained with spaCy's PoS and NER taggers (Honnibal and Johnson, 2015). To compare the distributions, we conducted statistical analysis on each parameter using the non-parametric Kruskal-Wallis test followed by the post-hoc Tukey test for pairwise comparisons. The analysis of PoS-tag distributions (Figure 3 (left)) does not show significant differences between matched and unmatched sentences from the article's main text. This suggests annotators did not exhibit a particular PoS preference when highlighting matched sentences. Yet, the POS-tag distribution of the main text sentences differs significantly from all kinds of captions. The details of the results are listed in Appendix Table 4. There are also significant differences between the captions types in terms of nouns, proper nouns, and determiners. The original captions are more distinct – they contain a noticeably higher proportion of proper names but a lower percentage of verbs, adverbs, and auxiliaries. The generated captions tend to have more nouns compared to human-generated captions. Just the opposite pattern is observed for the use of proper nouns. As expected, generated models avoid using this type and prefer generalized nouns. We observed no striking difference among the generated caption models except the clip-ft-gpt, which produces more proper nouns and fewer verbs.

The NER-tag analysis shows that human-generated wiki captions mostly contain entities that refer to a person, while generated captions avoid it. The IBM model's use of named entities is negligible in general. The details of the NER Distribution are presented in Figure 5 in the Appendix A.1.

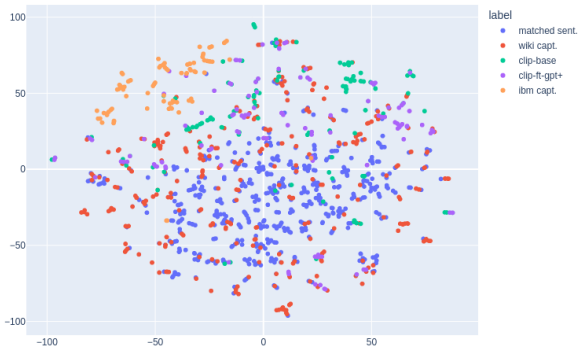


Figure 4: TSNE plot of SBERT sentence embeddings for matched sentences (blue), wiki captions (red), clip-base captions (green), clip-ft-gpt+ captions (purple) and IBM captions (orange)

To examine how sentences and captions are distributed in the semantic space, we plot their embeddings computed with SBERT, shown in Figure 4[7]. IBM-MAX captions cluster together and are located farther from the main text and ClipCap captions. Similarly, ClipCap captions are located in a specific area of the space, while original (wiki) captions, clip-ft-gpt+ captions, and matched sentences are distributed more widely. This corroborates the observation that captions show different linguistic properties and styles and sentences from the article's main text and, additionally, suggests that sentences may be more varied and linguistically diverse compared to generated captions.

5.2 Results: Intra-document Retrieval of Sentences and Images

We now compare different embedding models in terms of their ability to align sentences and captions, and sentences and images, using retrieval accuracies. We calculate the ranks of the target sentence, caption, or image (see Section 4) and report top-1 and top-5 accuracies. Additionally, the mean similarity scores between (un)matched sentences, captions, and images are presented in the Appendix Table 6 and Table 7.

The top-k accuracy scores for (i) sentence-to-caption/image and (ii) caption/image-to-sentence retrieval are presented in Table 2 and Table 3 respectively. Results from SBERT and MCSE are based on sentence-caption alignment, whereas CLIP and VILT results show sentence-image alignment. This allows us to compare unimodal to multimodal retrieval. We report retrieval accuracies on the paragraph-, text- and corpus level, as explained in Section 4.

In Table 2 and Table 3, we observe that the top-1 retrieval accuracy is overall very poor, even in the simpler match condition. On the paragraph level, the highest score for the matched sentences at top-1 is 0.66, achieved by multimodal retrieval with CLIP (in Table 2). The VILT model produces a slightly lower score, while the SBERT and MCSE models are notably low on aligning at paragraph level. For the article and corpus level, the top-1 accuracies are drastically low, in particular for caption/image-to-sentence alignment. Generally, caption/image-to-sentence retrieval is more complex than sentence-to-caption/image retrieval, regardless of the model.

[7]We use TSNE in scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

The top-5 accuracies look more promising across models and settings in the match condition, but it should also be noted that when retrieving from the paragraph-related sets the size of the candidate set is often less than five items. In the more realistic scenario of article-level retrieval, sentence embeddings (text-only and multimodal) perform better. The lowest retrieval accuracy is observed at the corpus level, as expected.

When we look at the retrieval scores of all sentences (column "all" in Table 2 and Table 3), the performance of SBERT and MCSE models further decreases, while average multimodal retrieval scores with CLIP and VILT is higher for all sentences than the matched sentences. This means that CLIP and VILT models will favor irrelevant images/sentences compared to relevant ones in top-1 and top-5 retrieval.

Finally, we look at the differences between various caption types and their similarities to the sentences. In sentence-to-caption conditions, for both SBERT and MCSE models, the generated captions are better at the paragraph and article level alignment. In contrast, the retrieval scored of wiki captions are higher at the entire set level. Among the generated captions, the clip-base model is a better fit for the task.

6 Discussion

We introduced a dataset for text–image alignment in multi-paragraph, multi-image documents, connecting captioned images with text spans from the main text which are depicted in the image. Our experiments show that these annotations provide a valuable benchmark dataset to evaluate the capabilities of zero-shot unimodal and multimodal pretrained models, that are challenged by image-text alignment in long and domain-specific documents. Based on the results, we revisit our research questions and possible implications of our experiments for future research on multimodal documents.

Do descriptions of images in articles show different linguistic properties than captions of the corresponding images? Yes. The analysis in Section 5.1 shows that descriptive, matched sentences from the main text exhibit different POS and NER distributions compared to the original captions written by Wikipedia authors. This highlights the importance of moving beyond the strong focus on captions in L&V research and indicates that different types of descriptions occurring within (and across) documents may exhibit different linguistic phenomena for visual language grounding.

Do the original captions in Wikipedia differ systematically from captions generated by captioning models? Partially. The analysis in Section 5.1 indicates that original captions written by Wikipedia authors differ in some aspects from the generated captions, which we expect to reflect the style of crowdsourced captions that many L&V models are currently trained on. This is not surprising but showcases that the style of captions collected in annotation and crowdsourcing experiments differs from naturally occurring captions found in real documents. This may bias or limit L&V models in a way that they do not encounter descriptive, visually grounded language in its full breadth in their pretraining data.

Are similarities between descriptive sentences within a text and captions robust enough to serve as a baseline for intra-document retrieval? Partially. The results in Section 5.2 show that intra-document retrieval for sentences and images via their captions works when the set of images/captions is restricted to the paragraph level, but drastically decreases at the article level. This holds for different types of captions. The retrieval score analysis shows inconclusive results in terms of the effect of captioning on different models.

How do image-sentence retrieval baselines compare to caption-sentence retrieval? The results in Section 5.2 show that sentence embeddings can distinguish more accurately between matched and unmatched sentences than multimodal models when looking at retrieval within an entire article. We believe that this may be because existing L&V models are typically trained on short texts that prioritize visually grounded language, but rarely on datasets of longer texts that include non-descriptive sentences. Generally, it appears that the multimodal models we tested lack awareness of depictability (i.e. detecting language that is visually grounded). Uni-modal sentence embedding models, on the other hand, seem to be less accurate in distinguishing grounded from non-grounded sentences at the more fine-grained paragraph level. For applications like intra-document retrieval in text-dominated documents, unimodal sentence embeddings still provide a better solution, but multimodal models have complementary strengths at the more fine-grained paragraph level distinctions. It

Table 2: Top-1 and Top-5 Retrieval Accuracy Scores for the sentence to caption/image conditions. The underlined scores represent the highest retrieval performance along the vertical axes. The match condition restricts candidate sentences to matched sentences.

		Top-1 paragraph		Top-1 article		Top-1 entire set		Top-5 paragraph		Top-5 article		Top-5 entire set	
	caption_type	**Match**	**All**	**Match**	**All**	**Match**	**All**	**Match**	**All**	**Match**	**All**	**Match**	**All**
SBERT	**wiki**	0.54	0.50	0.24	0.21	0.09	0.05	0.98	0.98	0.66	0.62	0.18	0.12
SBERT	**clip-base**	0.56	0.54	0.21	0.21	0.02	0.01	0.99	0.99	0.73	0.68	0.09	0.07
SBERT	**clip-ft-gpt+**	0.56	0.54	0.20	0.18	0.02	0.01	0.99	0.98	0.70	0.66	0.09	0.07
MCSE	**wiki**	0.52	0.50	0.20	0.19	0.05	0.03	0.99	0.99	0.68	0.63	0.11	0.08
MCSE	**clip-base**	0.58	0.53	0.23	0.19	0.01	0.01	1.00	0.99	0.71	0.68	0.09	0.06
MCSE	**clip-ft-gpt+**	0.55	0.53	0.19	0.18	0.01	0.01	0.99	0.99	0.68	0.67	0.08	0.06
CLIP	**wiki**	0.66	0.72	0.14	0.20	0.00	0.01	0.99	1.00	0.56	0.60	0.02	0.02
VILT	**wiki**	0.65	0.71	0.19	0.18	-	-	0.99	0.99	0.62	0.59	-	-

Table 3: Top-1 and Top-5 Retrieval Accuracy Scores for the caption/image to sentence conditions. The match condition restricts candidate sentences to matched sentences.

		Top-1 paragraph		Top-1 article		Top-1 entire set		Top-5 paragraph		Top-5 article		Top-5 entire set	
	caption_type	**Match**	**All**	**Match**	**All**	**Match**	**All**	**Match**	**All**	**Match**	**All**	**Match**	**All**
SBERT	**wiki**	0.24	0.14	0.08	0.04	0.04	0.02	0.85	0.73	0.25	0.17	0.09	0.05
SBERT	**clip-base**	0.22	0.15	0.04	0.03	0.00	0.00	0.83	0.73	0.15	0.13	0.01	0.01
SBERT	**clip-ft-gpt+**	0.21	0.15	0.04	0.03	0.00	0.00	0.83	0.73	0.17	0.12	0.02	0.00
MCSE	**wiki**	0.24	0.14	0.09	0.04	0.03	0.01	0.83	0.73	0.25	0.16	0.07	0.04
MCSE	**clip-base**	0.22	0.15	0.05	0.03	0.00	0.00	0.84	0.73	0.20	0.13	0.01	0.01
MCSE	**clip-ft-gpt+**	0.20	0.15	0.04	0.03	0.00	0.00	0.84	0.73	0.17	0.13	0.01	0.01
CLIP	**wiki**	0.16	0.16	0.01	0.03	-	-	0.76	0.75	0.09	0.13	-	-
VILT	**wiki**	0.15	0.16	0.01	0.03			0.76	0.75	0.09	0.13		

seems to be a promising direction for future work to explore models that exploit sentence-image and sentence-caption alignment in a joint fashion, and to develop multi-modal models that can handle text that includes non-descriptive language.

7 Conclusion

Wikipedia articles represent a genre of multimodal text that contains large amount of textual and visual information. Some foundational linguistic work on multimodal texts (Delin and Bateman, 2002; Hardy-Vallée, 2016) argues that in order to analyze multimodal texts, elements from different modalities should equally be treated as part of the document. With state-of-the-art L&V models being able to jointly represent text and image elements, this becomes increasingly feasible to do computationally as well. However, longer and more complex multimodal texts are not the norm in L&V research. With the collection of *WikiScenes with Descriptions*, we take a first step towards tackling the challenge of image-text alignment in naturally occurring, text-heavy, multi-image documents. This represents an important step in empirically-informed research on the topic of multimodal documents and provides a dataset for future modeling.

Limitations

Our extension of WikiScenes is a relatively small, domain-specific dataset so the results presented in this paper should not be assumed to necessarily generalize to other domains. The models used for the retrieval tasks were achieved with the respective base models and were not fine-tuned in our specific domain.

Ethics Statement

Images in the dataset are either under CC3.0 licenses or Open Domain. They are attributed via their identifications in Wikimedia Commons. We did not collect any personal information from annotators. Annotators were not presented with harmful materials during data collection. Crowdworkers were paid 0.35$ per item, which translates to an hourly wage of 9.01$.

Acknowledgement

The authors acknowledge financial support by the project "SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), funded by the program "Netzwerke 2021" of the Ministry of Culture and Science of the State of Northrhine Westphalia, Germany.

References

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. Cite: A corpus of image-text discourse relations. *arXiv preprint arXiv:1904.06286*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2019. An annotated dataset of coreference in english literature. *arXiv preprint arXiv:1912.01140*.

John Bateman. 2008. *Multimodality and genre: A foundation for the systematic analysis of multimodal documents*. Springer.

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.

Gullal S Cheema, Sherzod Hakimov, Eric Müller-Budack, Christian Otto, John A Bateman, and Ralph Ewerth. 2023. Understanding image-text relations and news values for multimodal news analysis. *Frontiers in Artificial Intelligence*, 6:1125533.

Judy Delin and John Bateman. 2002. Describing and critiquing multimodal documents. *Document Design*, 3:140–155.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Abbas Ghaddar and Philippe Langlais. 2016. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142.

Michel Hardy-Vallée. 2016. Text and image: a critical introduction to the visual/verbal divide by john a. bateman. *Visual Studies*, 31:366–368.

Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. 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 2034–2045, Hong Kong, China. Association for Computational Linguistics.

Laura Hollink, Adriatik Bedjeti, Martin Van Harmelen, and Desmond Elliott. 2016. A corpus of images and text in online news. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1377–1382.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1373–1378.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. 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 4622–4632, Hong Kong, China. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014a. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014b. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Fuxiao Liu, Hao Tan, and Chris Tensmeyer. 2023. Documentclip: Linking figures and main body text in reflowed documents. *arXiv preprint arXiv:2306.06306*.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Masayasu Muraoka, Ryosuke Kohita, and Etsuko Ishii. 2020. Image position prediction in multimodal documents. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4265–4274, Marseille, France. European Language Resources Association.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shigehiko Schamoni, Julian Hitschler, and Stefan Riezler. 2018. A dataset and reranking method for multimodal MT of user-generated image captions. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 140–153, Boston, MA. Association for Machine Translation in the Americas.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, online.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663.

Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, and Noah Snavely. 2021. Towers of Babel: Combining images, language, and 3D geometry for learning multimodal vision. In *ICCV*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014a. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014b. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Miaoran Zhang, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow. 2022. MCSE: Multimodal contrastive learning of sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5959–5969, Seattle, United States. Association for Computational Linguistics.

A Appendix

A.1 Text Analysis (Cont.)

Compared to the sentences, ClipCap captions contains similar amount of entities that refer to nationalities or religious or political groups (NORP), and significantly higher proportion of the dates or time periods (DATE). There is one named-entity category, ORGANIZATION was observed at similar rates among all textual elements.

Table 4: Statistical Difference between (i) matched and unmatched sentences and (ii) sentence, wiki captions and clip-ft-gpt20e captions in terms of POS- and NER-tag uses

	Sentence-Caption-Image
NOUN	554.19 (0.01 at all levels)
PROPN	105.79 (0.01 at all levels)
ADV	194.75 (0.01 sentence vs both captions)
VERB	765.87 (0.01 sentence vs both captions)
DET	494.13 (0.01 at all levels)
ADP	587.56 (0.01 at all levels)
AUX	636.11 (0.01 sentence vs both captions)
PERSON	84.92 (0.01 at all levels)
NORP	38.58 (0.01 at all levels)
DATE	120.86 (0.01 sentence vs both captions)
ORG	29.33 (0.01 clip-ft versus sent. and wiki capt.)

A.2 Annotation Instructions

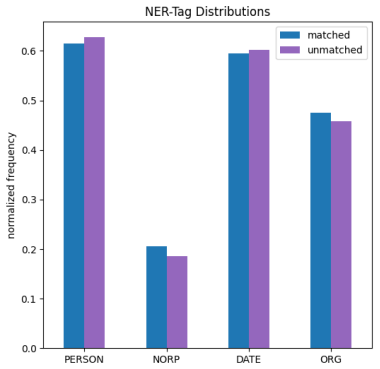
Figure 6 shows the annotation instructions used for collecting annotations that align/match text spans and images from crowd workers.

A.3 Computational Resources

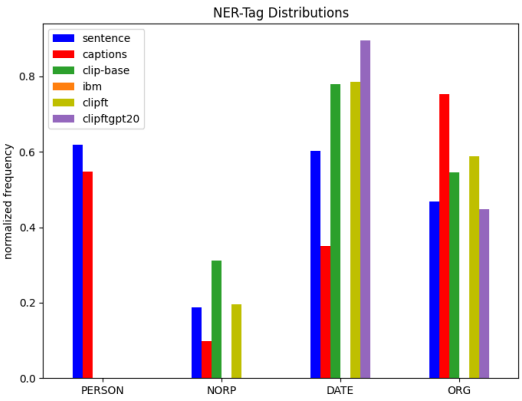
The experiments are conducted on a GPU workstation with NVIDIA® RTX™ A6000 (48GB). Table 5 list the approximate total time spent for ex-

Table 5: Analysis time (extracting embeddings and computing similarities) for each model on each condition

	sentence-to-caption/image	**image/caption-to-sentence**
SBERT	around 1 hours	3 hours
MCSE	around 2 hours	8 hours
CLIP	(all >32 hours) 4 hours[8]	10 hours
VILT	(all >2 days hours) 8 hours	19 hours



(a) NER tag distributions for matched and unmatched sentences



(b) NER tag distributions for matched sentences and captions

Figure 5: Comparing NER-tag distributions between textual elements

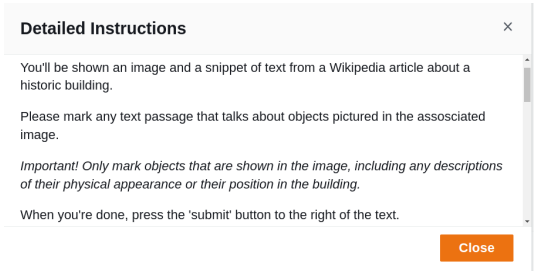
tracting the embeddings for each element (sentence, caption and image) and computing the similarities.

A.4 ClipCap finetuning

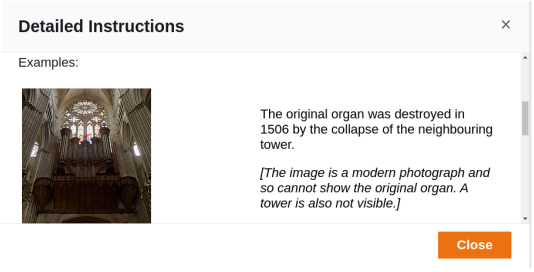
ClipCap finetuning follows the instructions from the original code repository: https://github.com/rmokady/CLIP_prefix_caption. First, the image is preprocessed using CLIP ("ViT-B/32") and mapped to a prefix vector. The prefix vector is projected into embedding space using a finetuned ClipCap model pretrained on Conceptual Captions. The prefix embedding is used as input for the GPT-2 model, as part of the ClipCap model. Greedy sampling with top-p=0.8 is used to generate the output sequence.

A.5 Similarity based Analysis

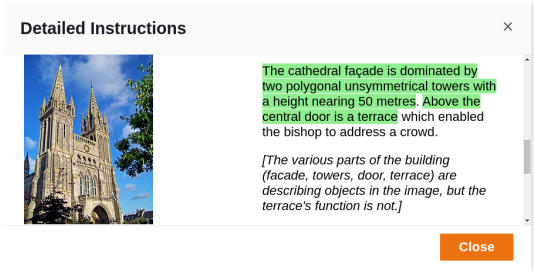
Table 6 and Table 7 present the average similarity scores of the target item against various candidate sets in two directions; sentence-to-caption/image and caption/image-to-sentence respectively.



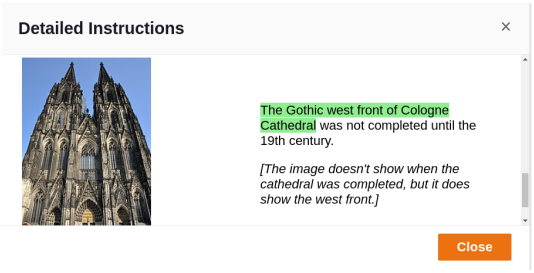
(a) Main instructions for paragraph-image alignment annotation



(b) Example shown in instructions for paragraph-image alignment annotation with none of the text spans matching the visual content of the image



(c) Example shown in instructions for paragraph-image alignment annotation with a longer text span matching the visual content of the image



(d) Example shown in instructions for paragraph-image alignment annotation with a shorter text span matching the visual content of the image

Figure 6: Instructions used for the collection of annotations on paragraph-image alignments

Table 6: Average similarity scores for the sentence-to-caption or sentence-to-image conditions. Bold face represents the highest score along the horizontal axes, while the underlined text corresponds to highest score among the three caption types within each embedding space.

		paired		paragraph		article		entire	
		Unmatched	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched
SBERT	**wiki**	0.754	**0.767**	0.754	**0.767**	0.745	0.752	0.733	0.738
SBERT	**clip-base**	0.744	**0.752**	0.743	0.751	0.739	0.746	0.726	0.731
SBERT	**clip-ft-gpt+**	0.752	**0.764**	0.752	0.762	0.750	0.759	0.739	0.746
MCSE	**wiki**	0.174	**0.216**	0.176	0.212	0.154	0.179	0.122	0.140
MCSE	**clip-base**	0.187	**0.221**	0.187	0.216	0.180	0.206	0.146	0.167
MCSE	**clip-ft-gpt+**	0.202	**0.235**	0.203	0.232	0.198	0.226	0.167	0.193
clip	**wiki**	**0.813**	0.772	0.810	0.780	0.803	0.785	0.791	0.775

Table 7: Average similarity scores for the caption-to-sentence or image-to-sentence conditions

		paired		paragraph		article		entire	
		Unmatched	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched
SBERT	**wiki**	0.754	**0.767**	0.756	0.756	0.746	0.743	0.736	0.735
SBERT	**clip-base**	0.743	**0.751**	0.745	0.744	0.741	0.739	0.737	0.737
SBERT	**clip-ft-gpt+**	0.752	**0.764**	0.754	0.756	0.752	0.752	0.748	0.748
MCSE	**wiki**	0.174	**0.216**	0.180	0.186	0.158	0.154	0.135	0.132
MCSE	**clip-base**	0.187	**0.221**	0.192	0.196	0.184	0.182	0.175	0.175
MCSE	**clip-ft-gpt+**	0.202	**0.235**	0.207	0.213	0.204	0.200	0.192	0.189