Assessing News Thumbnail Representativeness: Counterfactual text can enhance the cross-modal matching ability

Yejun Yoon[†] **Seunghyun Yoon**[‡] **Kunwoo Park**[†⋆]
[†]Department of Intelligent Semiconductors, Soongsil University
[‡]Adobe Research, USA
[⋆]School of AI Convergence, Soongsil University
yeayen789@gmail.com, syoon@adobe.com, kunwoo.park@ssu.ac.kr

Abstract

This paper addresses the critical challenge of assessing the representativeness of news thumbnail images, which often serve as the first visual engagement for readers when an article is disseminated on social media. We focus on whether a news image represents the actors discussed in the news text. To serve the challenge, we introduce NEWSTT, a manually annotated dataset of 1000 news thumbnail images and text pairs. We found that the pretrained vision and language models, such as BLIP-2, struggle with this task. Since news subjects frequently involve named entities or proper nouns, the pretrained models could have a limited capability to match news actors' visual and textual appearances. We hypothesize that learning to contrast news text with its counterfactual, of which named entities are replaced, can enhance the cross-modal matching ability of vision and language models. We propose CFT-CLIP, a contrastive learning framework that updates vision and language bi-encoders according to the hypothesis. We found that our simple method can boost the performance for assessing news thumbnail representativeness, supporting our assumption. Code and data can be accessed at https://github.com/ssu-humane/news-images-acl24.

1 Introduction

This study investigates the representativeness of news thumbnails, which are images displayed as previews of news articles. Visual content often carries a more persuasive impact and leaves a long-lasting impression compared to text (Joffe, 2008; Newman et al., 2012; Seo, 2020). Consequently, the misrepresentation in news thumbnails can cause more critical consequences. Despite its importance, there have been a few studies addressing the representativeness of news images (Hessel et al., 2021; Choi et al., 2022). Achieving the goal requires a model to understand the cross-modal relationship between visual and textual content.

Figure 1: An illustration of the key idea of the proposed method. To assess whether a news thumbnail image represents the body text, the method generates counterfactual text to be used as negative samples for contrastive updates.

Researchers in Natural Language Processing and Computer Vision communities have explored multimodal technologies by addressing general vision and language (V+L) tasks, including visual question answering (Antol et al., 2015), image-text retrieval (Cao et al., 2022), and visual entailment (Xie et al., 2019). Recent approaches have leveraged contrastive learning techniques for representing distinct modalities in the same vector space. CLIP (Radford et al., 2021) stands out as one of the pioneering methods that achieve remarkable performance, aimed at increasing the similarity between image-text pairs (referred to as *positive* samples) compared to non-paired image and text (referred to as *negative* samples). While contrastive learning was not a new concept (Oord et al., 2018), CLIP outperformed preexisting methods by utilizing a vast collection of web-based image and text data for pretraining. Recent advancements have further enhanced the performance by refining learning objectives (Li et al., 2022b), incorporating cross-attention layers into the model architecture (Li et al., 2021a), and augmenting the training data with high-quality samples (Li et al., 2022a; Hao

Findings of the Association for Computational Linguistics ACL 2024, pages 9009–9024

et al., 2023).

Despite the progress made in multimodal methods for general vision-and-language benchmarks, we hypothesize that evaluating the representativeness of thumbnails remains difficult, even for the most advanced models, due to two primary challenges. **(1) Preference for general descriptions in popular benchmarks**: General V+L datasets, such as Conceptual Captions (Sharma et al., 2018) or COCO-Caption (Chen et al., 2015), were primarily designed to enhance a model's ability to comprehend general behaviors or scenes. Consequently, these datasets deliberately avoid specific descriptions about entities (e.g., "The man at bat readies to swing at the pitch"). In contrast, news events typically revolve around political actors or organizations, leading to news texts that frequently reference named entities (Luo et al., 2021). **(2) Lack of dataset labeled according to an objective definition**: Representativeness is an inherently abstract and subjective concept, where even humans may disagree with each other. Developing an objective definition is imperative to obtain high-quality labels that different human annotators can agree on and facilitate the development of proper technologies. While there was a human-labeled dataset (Choi et al., 2022), they asked for subjective opinions of human annotators without providing objective definitions grounded by journalism principles.

To address (1) and (2), borrowing the principle of five Ws in journalism (The Associate Press, 2022), we introduce the evaluation dataset of NEWS Thumbnails and Text pairs, NEWSTT. The annotators were instructed to label the 1000 pairs through the lens of *Who*, on whether an image represents news actors. Using the high-quality labels reflecting the objective definition of representativeness, we evaluated the zero-shot ability of vision and language models to understand news actors' visual portrayals. We found that CLIP performed better at the task than more recent methods, such as BLIP-2, suggesting the efficiencies of its dual encoder architecture and contrastive objective. To improve its matching ability, this study introduces CFT-CLIP, a CounterFactual Text-guided Contrastive Language-Image Pretraining framework. As illustrated in Figure 1, the key idea of the proposed framework is to generate counterfactual news text where the actor of a news event is changed to be used as hard negatives for contrastive learning. In the news text example, the actor is identified as 'Biden,' and its corresponding token is replaced with 'Trump' by the counterfactual text generator. Thus, the generated sentence no longer represents the key actor of the news event depicted in the thumbnail image. Our research hypothesis is that *learning to contrast the generated counterfactual text with the original news text will enhance the ability of vision language models to assess the representativeness of a thumbnail image*.

We summarize the contributions of this study.

1. We address the problem of assessing news thumbnail representativeness by determining whether a given news image depicts the actors of the news events (Who) according to the principle of 5Ws.

2. To serve the task, we introduce NEWSTT, a manually annotated dataset of 1,000 news thumbnails and text pairs with high-quality labels. We found that general vision language models struggle with it.

3. We propose CFT-CLIP, a counterfactual text-guided contrastive language-image pretraining framework. We found that our simple method could outperform larger pretrained and domain-adapted models, supporting the research hypothesis.

2 Related Works

2.1 Vision language contrastive pretraining

Researchers have worked on pretraining methods to tackle V+L tasks such as image-text retrieval. Lu et al. (2019) extended BERT to a multi-modal model by co-attentional transformer layers. Another study proposed to learn a universal image and text representation by four pre-training tasks, including masked language modeling and image-text matching (Chen et al., 2020). Researchers proposed CLIP, a training scheme of visual representation by learning directly from paired text using contrastive loss (Radford et al., 2021). Being trained on a web-scaled dataset, the transformer-based visual representation achieved state-of-the-art results in zero-shot tasks. Another study showed its potential as a reference-free image-captioning metric (Hessel et al., 2021). Other researchers presented ALIGN, a bi-encoder trained by a contrastive loss similar to CLIP (Jia et al., 2021). While we used CLIP encoders as the target backbone in this study, our contrastive learning framework can be applied to

a more recent vision-and-language model that incorporates the image-text alignment, such as BLIP-2 (Li et al., 2023).

Research has underscored the significance of incorporating hard negative samples into the contrastive objective (Nishikawa et al., 2022; Robinson et al., 2021). In response, we introduce a method to generate hard negative text samples using a masked language model. Although a handful of studies have explored the use of masked language models for generating hard negatives (Li et al., 2021b; Yao et al., 2022), our approach stands apart due to its distinct token selection and prediction procedures in the step of counterfactual text generation. We compared its effectiveness against the autoregressive generation of GPT-3.5 Turbo in the ablation experiments.

2.2 Multimodal misinformation

This study falls into the broad category of research on multimodal misinformation. Previous research mainly worked on fact verification involving multimodal claims or evidence (Zlatkova et al., 2019; Shu et al., 2020; Yao et al., 2023; Rani et al., 2023). In contrast, this study examines whether news thumbnails and text present the same actor. A relevant problem is image repurposing (Luo et al., 2021; Abdelnabi et al., 2022), where a *real* image is used out-of-context to create realistically looking misinformation. To tackle the problem, Luo et al. (2021) proposed NewsCLIPPings, a dataset of out-of-context image detection in the news context. They created the mismatched image-text pairs by swapping the matches of the collected news image and text pairs. Similarly, Mishra et al. (2022) created a dataset of multimodal fact verification, of which false claims were obtained by manipulating the pairs. Unlike the previous research on out-of-context image detection, this study aims to focus on whether news subjects are represented in the thumbnail image. We also created a manually annotated evaluation dataset, contrasting with previous datasets based on synthetic labels (Luo et al., 2021; Mishra et al., 2022). Our target task is related to but differs from the manipulated image detection (Huh et al., 2018; Rossler et al., 2019), known as deepfakes or cheapfakes, which aimed to detect a *fake* image generated by AI technologies or simple tools.

3 Target Problem: Assessing News Thumbnail Representativeness

We aim to address the problem of automatically evaluating the representativeness of a news thumbnail image for its article. The task is critical due to the two reasons. First, the sharing of news online is abundant (Park et al., 2021a; Lee and Ma, 2012; Hermida et al., 2012). Second, the thumbnail often serves as the initial and sole visual interaction for social media users when a news link is shared. Visual content is perceived as more credible and leaves a long-lasting impression than text (Joffe, 2008; Seo, 2020; Newman and Zhang, 2020). Given these considerations, thumbnail images that do not accurately represent their content can misguide the reader's understanding in online environments.

We formally define the problem as a binary classification. Given a news thumbnail I and its news text T, we aim to develop a classifier $f_\theta(I, T)$ predicting a binary label L on predicting whether I represents T. We assign 1 to L if I represents T; otherwise, 0. I is deemed representative of T if I portrays at least one of the actors of a news event, which can be identified from T. This study aims to develop a vision language model that predicts L from a pair of I and T. We assume the task in a zero-shot setting, where the labeled data is limited. The reference news text T can be either a title or a summary.

4 NEWSTT: A Dataset of Thumbnail Representativeness for News Text

4.1 Raw data collection

NELA-GT-2021 is a headline-oriented[1] news corpus that encompasses nearly all news articles published by 367 news sources throughout the year 2021 (Gruppi et al., 2022). Since the dataset does not provide web links to thumbnail images, we conducted a supplementary data collection step. By referencing the URLs of news articles available in the dataset, we parsed the HTML document and extracted the thumbnail image URL from the meta tag with the og:image attribute. As a result, we obtained 442,741 pairs of news headlines and thumbnail images, which were sourced from 81 different news media outlets. According to the media ratings provided by Media Bias/Fact

[1]They put the body text whose random tokens are masked.

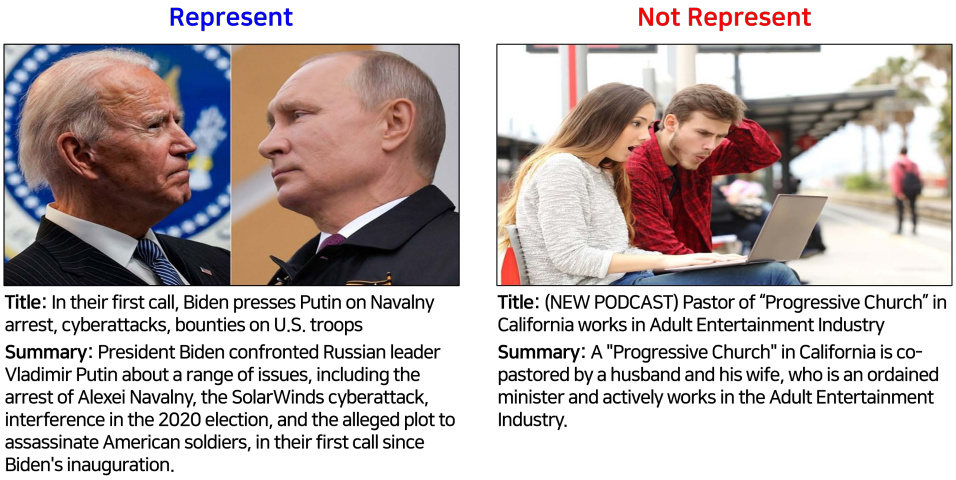

Figure 2: Labeled data examples.

Check (MBFC)[2], the datasets were well-balanced for the political orientation and trustworthiness ratings. Subsequently, we partitioned this unlabeled dataset into 427,741/5,000/10,000 pairs for train/validation/test purposes while keeping the distribution of the number of articles over different media sources. The validation split was reserved for hyperparameter optimization, and the test split served as the source for labeled data construction.

4.2 Data annotation

We describe the process for building the annotation guideline. Following a previous study (Choi et al., 2022), we conducted the pilot labeling task given an abstract definition of a representative thumbnail, which is "*an image that visually conveys the news event that can be identified from the accompanying text*". There were often disagreements in the resulting labels likely due to the subjective nature of definition. Following internal discussions and consultation with a journalism expert with a Ph.D. in Communication, we focused on the key dimension of news events corresponding to the five Ws (The Associate Press, 2022): on whether the image represents news subjects, i.e., *Who*.

For the annotation process, we sampled 1,000 news articles from the pool of 10,000 instances of the test split. Relying on the MBFC ratings, we sampled 500 articles from the trustworthy outlets, labeled as *mixed* or *high* for the trustworthiness rating, and the remaining 500 from the fake news sources, labeled as *low*. We manually collected the body text for the 1000 articles and ran the OpenAI API for obtaining the summary text by GPT-3.5 Turbo[3]. We hired three English-speaking students who frequently read news online from one of the authors' institutions. Given a news thumbnail, title, and summary text, the annotators were asked to answer whether the news actors identified in the news text are visually portrayed in the thumbnail image. To facilitate the annotation process, we instructed them to identify the actors of the news event by referring to the text and, in turn, to find visually depicted ones. Detailed annotation guidelines can be found in Appendix F. The annotation results demonstrated a high inter-annotator agreement rate, with a Krippendorff's alpha of 0.705, ensuring the quality and reliability of the labels.

	Represent (1)		Not represent (0)	
	Title	Summary	Title	Summary
Words	14.9	39.1	14.7	39
Nouns	3.45	8.63	3.75	9.7
Verbs	1.81	4.06	1.68	4.09
Adjectives	0.93	2.39	1.19	2.73
Named entities	2	3.41	1.64	2.76

Table 1: Mean counts of words, part-of-speech units, and named entities measured on the labeled text.

4.3 Data analysis

NEWSTT consists of 817 representative and 183 non-representative image-text pairs. Table 1 presents the descriptive characteristics of the dataset for each class. While the samples in the two classes are similarly long, the news text of label 1 tends to include more named entities. Figure 2 presents an example for each class. The label 1 ex-

[2] https://mediabiasfactcheck.com/

[3] The used prompt is in Appendix E.

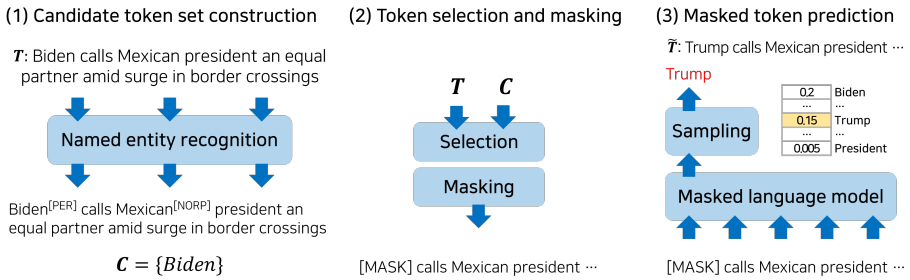

Figure 3: An illustration of the counterfactual text generation process by CFT-CLIP. We select named entity tokens in an original text that could indicate news subjects. A masked language model generates a counterfactual text by predicting new tokens for the selected entity tokens.

ample shows the image portraying Biden and Putin, who are the actors of a news event. The thumbnail image on the right does not show the news actor but arouses the potential reader's interest by showing the surprised faces.

5 Methods

To tackle the task of assessing thumbnail representativeness in a zero-shot setting where no labeled data is available for training, we use a thresholding classifier based on the cross-modal similarity,

$$f(I,T) = \mathbb{1}(sim(\mathbf{v}_I, \mathbf{v}_T) > \tau), \quad (1)$$

where τ is a thresholding hyperparameter, $sim(\cdot)$ is cosine similarity, $\mathbb{1}(\cdot)$ is an indicator function, and $\mathbf{v}_I$ and $\mathbf{v}_T$ indicate the feature embedding of I and T, respectively. Assuming the parity prior, we set τ as the median of $sim(\cdot)$ on the validation set.

5.1 Background: CLIP

CLIP is a vision and language bi-encoder that represents each modality in a fixed-dimensional vector space (Radford et al., 2021). The model is based on a bi-encoder architecture, $f_{image}(\cdot)$ and $f_{text}(\cdot)$. For an input image I and text T, the two encoders return the image embedding $\mathbf{v}_I$ and the text embedding $\mathbf{v}_T$, respectively. The parameters were trained via contrastive learning on a large web collection of image-text pairs. Since $\mathbf{v}_I$ and $\mathbf{v}_T$ are in the same vector space, CLIP can serve as a backbone for the zero-shot thresholding classifier in Eq. 1.

5.2 Proposed method: CFT-CLIP

We propose CFT-CLIP, a counterfactual text-guided contrastive language-image pretraining method. The proposed framework aims to improve the vision and language bi-encoder by contrastive updates involving the counterfactual text generated from an input text. Figure 3 shows an example of generation. While the original news pertains to an event involving Biden's summit with the Mexican president, the generated sentence implies a counterfactual event centered on Trump by substituting the token 'Biden'. This substitution is accomplished via the prediction of a masked language model targeting the named entity token. Based on the news characteristics on the prevalent coverage of named entities (Park et al., 2021b; Müller-Budack et al., 2020), we hypothesize that contrasting the original news text with its counterfactual would make the vision and language bi-encoder to obtain a more effective representation for understanding a thumbnail image's representativeness.

5.2.1 Counterfactual text generation

Given a pair of image I and text T, we aim to generate a counterfactual news text $\tilde{T}$, which is semantically distinct from the anchor image I (and the original text T).

(1) Candidate token set construction: The first step aims to select tokens that likely refer to the actor of a news event. A part-of-speech tagger and named entity recognizer are applied to T, which returns a candidate token set C. We tested several strategies targeting named entity categories that are frequently used in news articles, such as person, organization, and GPE[4].

(2) Token selection and masking: We select tokens from C and mask them in the original text T using the special token [MASK]. To preserve the original context of T, we ensure that no more than 30% of its total tokens are masked. If the number of target tokens selected from C exceeds this threshold, we randomly select a subset of these tokens to meet the condition. Otherwise, all tokens in C are masked in T.

[4]Countries, cities, states

(3) Masked token prediction: Using a masked language model, e.g., BERT (Devlin et al., 2019), we predict new tokens for the masked positions in T. To avoid reconstructing the original text, we divide the logit scores of the output vector by a temperature parameter and repeatedly sample from the softmax-normalized distribution until a token different from the original is generated. This step produces $\tilde{T}$, which represents a synthetic event wherein the subject is modified from the original news event, yet it preserves the overarching topic of T. Thus, $\tilde{T}$ can be deemed a hard negative sample in contrastive learning, which diverges from I but remains more closely aligned with T than a random text or in-batch negatives.

In Section 6.3, we present the ablation experiments to show the effectiveness of the proposed counterfactual text generation. Additionally, we demonstrate several counterfactual text examples generated by the proposed method in Appendix D.2.

5.2.2 Training objective

We train an image and text bi-encoder model to minimize

$$-log\frac{e^{sim(\mathbf{v}_{I_i},\mathbf{v}_{T_i})/\tau}}{\sum_{j=1}^{N}\{e^{sim(\mathbf{v}_{I_i},\mathbf{v}_{T_j})/\tau}+e^{sim(\mathbf{v}_{I_i},\mathbf{v}_{\tilde{T}_j})/\tau}\}}, \quad (2)$$

where $\mathbf{v}_{I_i}$ is the feature embedding of the vision encoder for I_i, $\mathbf{v}_{T_i}$ is the feature embedding of the text encoder for T_i, $sim(\cdot)$ is the dot product, and N is the mini-batch size.

Minimizing Eq. 2 aligns the text representation with the image representation by increasing the similarity of positive pair (I_i, T_i) compared to those of negative pairs (I_i, T_j) and $(I_i, \tilde{T}_i)$. Since $\tilde{T}$ represents a distinct news event of which actors were replaced from T, the objective would help a bi-encoder model learn to capture whether the thumbnail image represents the news actors that can be identified from T.

5.2.3 Model architecture

Figure 4 depicts the neural architecture used for the proposed method. We initialized the model by adopting a pretrained CLIP checkpoint[5]. The image encoder f_{image} is the ViT-L/14 vision transformer (24 layers), and the text encoder f_{text} is a causal text transformer (12 layers). The image and text

[5] https://huggingface.co/openai/clip-vit-large-patch14

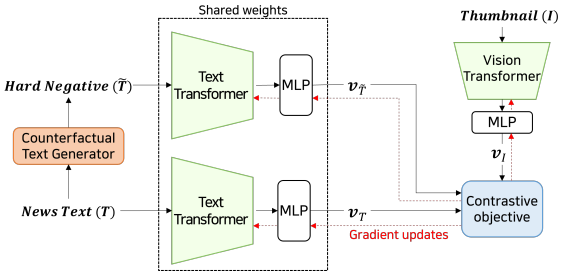

Figure 4: Neural architecture of CFT-CLIP

encoder are followed by an MLP pooler mapping the input to a 768-dimensional vector, respectively. T and $\tilde{T}$ are fed into the shared-weight text encoder, separately. According to the hyperparameter optimization experiment, we froze 11 out of 12 text transformer layers during training. All vision layers were updated.

5.2.4 Pretraining corpus

We used two pretraining corpora. First, we used the training split of NELA unlabeled data. Inspired by a prior study that replicated the CLIP pretraining based on a web data collection (Schuhmann et al., 2022), we applied the data sanitization process to the training split of unlabeled data by filtering only those with a cosine similarity of 0.28 or higher over CLIP embeddings. The resulting 105,737 pairs for the unlabeled training have a high likelihood of a representative thumbnail ($L = 1$). Since the body text in the NELA corpus is incomplete, we relied on the news title as the reference text for pretraining. Second, inspired by the recent research on the importance of high-quality samples for pretraining (Zhou et al., 2024; Schuhmann et al., 2022), we used a high-quality news corpus published by the BBC, a UK-based news source rated as trustworthy by MBFC (Verma et al., 2023). This corpus comprises 196,538 pairs, each containing a news title, an editor-written summary, and a thumbnail image. The corpus encompasses news articles released through the official website from 2009 to 2021. We applied the sanitation process to the dataset in the same manner as with the NELA corpus, leaving 58,210 samples for contrastive training. We ensured no overlap between the BBC training set and the NELA validation/test splits. Since NewsTT originates from the NELA corpus, we used the NELA validation split for hyperparameter optimization. In the experiments using the BBC corpus, we tested two variants using the news title and human-written summary as reference text.

6 Evaluation

We conducted evaluation experiments to understand the effects of the proposed method and test the research hypothesis. In Section 6.2, we examined how the pretrained and proposed vision language models perform for the target task. In ablation experiments (Section 6.3), we analyzed the effects of each module of the proposed method in more detail. Lastly, in Section 6.4, we conducted qualitative analyses to identify the error patterns where the proposed method fails.

6.1 Experimental setups

As baselines, we measured the ability of pretrained vision language models. In addition to CLIP, we used BLIP and BLIP-2, which are a family of vision language encoders that is based on bootstrapped training (Li et al., 2022a, 2023). They achieved state-of-the-art performance in the vision-language benchmarks. BLIP-2+SBERT is a pipelined approach that integrates BLIP-2 with SentenceBERT. By obtaining a caption text from an input image I by BLIP-2, we assessed its semantic similarity to the reference text using SentenceBERT. To investigate the effects of domain-adaptive pretraining, we also included CLIPAdapt as a baseline model, which was trained by using the CLIP objective on the training corpus. For inference, we used the summary text as the reference text T according to the results of an ablation experiment (Table A2).

To evaluate the baseline and proposed methods, we employed the f1 score and the Spearman rank correlation coefficient to evaluate binary prediction and cosine similarity scores, respectively. All experiments were conducted using five different random seeds, and we reported the average performance along with the standard error. The t-test was used for estimating statistical significance. We used the SpaCy pipeline for named entity recognition. A BERT-base checkpoint was used for the masked language prediction. All checkpoints used for experiments and implementation details can be found in Appendix B. We optimized all hyperparameters using the NELA validation split.

6.2 Main results

Table 2 presents the evaluation results of the baseline and proposed methods. Here, we reported the performance of CFT-CLIP and CLIPAdapt models that were pretrained using the BBC dataset with the summary text. For CFT-CLIP, we used the model targeting person-labeled entity tokens. The decisions were based on the ablation experiments on the effect of pretraining corpus (Table 5) and counterfactual text generation (Table 3), respectively. We made three observations. First, among the pre-trained models (the bottom four rows), CLIP achieved the best f1 of 0.763, and BLIP-2 was the best by the Spearman coefficient with 0.415. Given that BLIP-2 (1.17B) is 2.74 times larger than CLIP (427M), this observation suggests the effectiveness of CLIP's bi-encoder architecture for the target task. Second, CLIPAdapt outperformed all the pretrained models. This suggests that domain-adapted continued pretraining can improve the performance for the target task, congruent with the finding of a previous study (Gururangan et al., 2020). Third, CFT-CLIP outperformed all baseline methods with the f1 of 0.815 and the Spearman coefficient of 0.491 ($p<0.001$). The performance gap with CLIPAdapt, the second-best method, is significant: 0.048 for f1 and 0.032 for the Spearman coefficient. The finding suggests that incorporating the counterfactual text in its contrastive objective as hard negatives can improve the cross-modal matching ability of the vision language bi-encoder, supporting the research hypothesis.

Model	F1	Spearman
CFT-CLIP	**0.815**±0.003	**0.491**±0.005
CLIPAdapt	0.767±0.006	0.459±0.004
CLIP	0.763	0.409
BLIP	0.737	0.408
BLIP-2	0.707	0.415
BLIP-2 + SBERT	0.694	0.341

Table 2: Model comparison results

Target token	F1	Spearman
Person	**0.815**±0.003	**0.491**±0.005
Organization	0.784±0.002	0.443±0.002
GPE	0.762±0.004	0.410±0.005
All	0.785±0.003	0.457±0.003
Random (15%)	0.715±0.013	0.463±0.005
Random (30%)	0.68±0.007	0.461±0.002

Table 3: Varying performance by token selection strategies in counterfactual text generation

6.3 Ablation experiments

What tokens should be targeted? We evaluated the performance of CFT-CLIP variants aiming to replace different types of entity tokens in

counterfactual text generation. In particular, we targeted the frequently used named entity tokens in the news text: person, organization, and GPE. We also tested the model that replaces all three token types, denoted All in the table. The top 4 rows in Table 3 present the comparison results. We found that targeting person-labeled entity tokens achieved the best performance. This finding could be explained by the prevalent coverage of person entities, such as politicians, in news events (Park et al., 2021b; Müller-Budack et al., 2020). Additionally, we tested baseline strategies that replace tokens at random positions for being used for negative samples in contrastive learning, as investigated in previous studies (Nishikawa et al., 2022; Robinson et al., 2021). We evaluated two variants that select 15% and 30% tokens, respectively, which were predicted by the same BERT backbone as the proposed method. The bottom two rows in Table 3 show the model performance, achieving significantly worse f1 and Spearman coefficients than the CFT-CLIP variants ($p<0.001$). According to the results, we targeted person-labeled tokens in the proposed method.

Is the masked LM necessary? CFT-CLIP used a masked language model to generate counterfactual text for contrastive updates. An autoregressive large language model, such as GPT (Radford et al., 2018), could be used alternatively. To validate the idea, we ran the OpenAI API for generating counterfactual text for the BBC news summary by GPT 3.5-Turbo. The generated text was used for contrastive update, following the objective in Eq. 2. The used prompt is available in Appendix E. As shown in Table 4, the GPT-based contrastive update was not as successful as the proposed method based on a masked language model. The proposed select-and-replace approach could generate a more suitable counterfactual sample to be used for hard negatives, rather than rewriting the whole sentence conditioned on the original text.

We also tested a baseline approach that ablates a generation model for predicting masked tokens by randomly sampling person-labeled entities from the tokens from the training set. The simple method achieved an f1 of 0.77 and a Spearman coefficient of 0.455, respectively. This finding supports the effectiveness and necessity of counterfactual text generation in the proposed method.

Data quality vs. quantity Table 5 presents the performance of CFT-CLIP models with varying pretraining corpus. The BBC dataset resulted in superior model performance compared to the NELA dataset, following the same distribution of the labeled dataset. Given that the BBC is recognized for upholding high journalistic standards, this observation may indicate that leveraging a high-quality singular source is more beneficial than employing articles published by diverse news sources for contrastive pretraining. This finding is aligned with the recent research on large language models (Zhou et al., 2024), emphasizing that ensuring data quality is more important than merely increasing the size of the training corpus. In the experiments comparing title and editor-written summaries using BBC, we could not find a clear winner. While using the summary text led to the best f1 of 0.815, using the title achieved the best Spearman coefficient of 0.504. This suggests that both news headlines and summary text can serve as a useful proxy for news content, in line with established journalism principles (The Associate Press, 2022). Since f1 is a more proper metric for evaluating classification ability, we used BBC with the summary text for the other experiments.

Model	F1	Spearman
CFT-CLIP	**0.815**±0.003	**0.491**±0.005
GPT-based	0.640±0.018	0.445±0.002

Table 4: Comparison with the model trained with GPT-based counterfactual text

Data	F1	Spearman
BBC (T: summary)	**0.815**±0.003	0.491±0.005
BBC (T: title)	0.790±0.007	**0.504**±0.001
NELA (T: title)	0.772±0.003	0.448±0.002

Table 5: Varying performance by the pretraining corpus

6.4 Error analysis

To identify the remaining challenges, we analyzed error categories for the sampled 100 error cases. The first author identified the initial category by thematic coding, which were improved by the iteration of discussion with the other authors and re-annotation. We observed four recurring categories: entity recognition failures (46%), external knowledge required (14%), deep textual understanding required (14%), and deep visual understanding required (14%). Figure 5 presents two error examples. In the first example, which belongs to the entity recognition failure case, the model made a

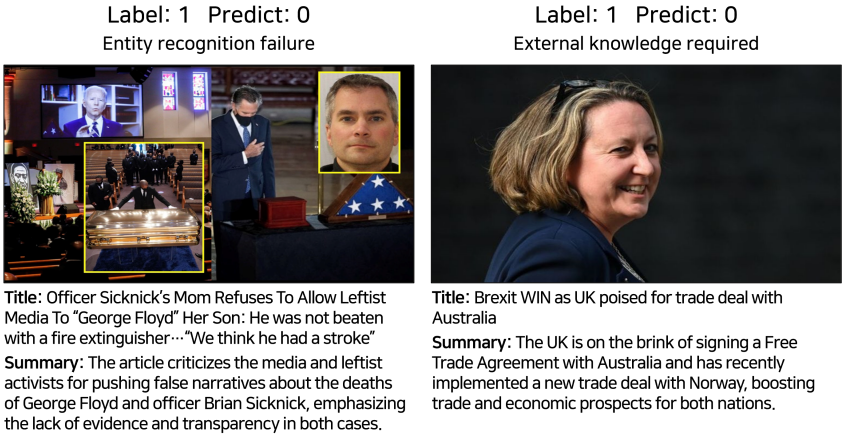

Figure 5: Error examples

wrong prediction possibly due to the failed identification of Brian Sicknick, the officer involved in the George Floyd case. In the second example, while the image shows Anne-Marie Trevelyan, a UK minister, the model could not find the cross-modal link without referring to external knowledge.

7 Discussion and Conclusion

Automatically evaluating a news thumbnail image's representativeness is important for holding journalistic standards and building a trustworthy online environment. To address the important but underexplored problem in the research community, this study introduced NewsTT, a paired dataset of news thumbnail images and text with high-quality labels on whether the image represents the actor of the news event. We investigated the use of vision language models for the zero-shot assessment. Our proposed CFT-CLIP outperformed larger pretrained vision language models and domain-adapted methods. This supports the research hypothesis that counterfactual news text, of which named entities are replaced by a masked language model, could enhance the cross-modal matching ability by contrastive learning.

In ablation experiments, we found that using the counterfactual text generated by an autoregressive transformer language model, GPT-3.5 Turbo, as negative samples could not achieve a better outcome than CFT-CLIP. While this supports the effectiveness of the proposed method, we do not conclude that masked language models are clear winners for the counterfactual news generation over autoregressive language models. After manually examining the results, we observed that GPT could produce plausible counterfactual news. By contrast, the generation by a masked language model sometimes led to imperfect generations, including broken grammar (Table A4). As shown in a recent study on text embedding (Lee et al., 2024), an additional step might be required to ensure the quality of the generated text.

This study has several future directions. First, according to the rule of 5Ws, this study focused on the *Who* aspect of news thumbnails. Future research could extend its focus to cover other aspects of 5Ws, such as whether the image represents the subject's action, i.e., *What*. Question-answering approaches could be investigated, as done in a recent study (Rani et al., 2023). Second, the proposed method could be improved to address the challenges identified in the error analysis (Section 6.4). Future studies could investigate the use of multimodal language models, such as InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2024), or GPT-4V (OpenAI, 2023), to allow for handling the errors involving *deep visual understanding*. The proposed CFT-CLIP could assist them with aligning the vision and language representation by adopting the counterfactual text. Third, this study could be extended to the broader research on computational social science. As similarly done in previous studies (Lommatzsch et al., 2022; Oostdijk et al., 2020), one could extend the target problem by addressing it as a ranking task rather than binary classification. It could enable a news application that recommends a suitable thumbnail image for a given news article. Additionally, the proposed CFT-CLIP could be adopted for automated fact verification involving image evidence (Luo et al., 2021; Mishra et al., 2022; Yao et al., 2023).

Limitations

This study bears several limitations. First, the scale of NewsTT is limited. Since the annotation task is complicated, we chose to do an in-lab annotation to provide high-quality labels for evaluation. Since the dataset reflects the distribution of real-world news articles and the label quality is high, it can be used for a reliable evaluation corpus. Future studies could scale up the dataset via crowdsourcing with the annotation scheme developed in this study. Second, since we continued to pretrain the CLIP encoders, CFT-CLIP inherits the weakness of the CLIP vision encoder. CLIP might be culturally biased toward the Western countries where the pretraining dataset may originate. The pretrained CLIP checkpoint used in this study cannot handle the entire body text as input because the maximum token length is 77, which is smaller than the average body text length. While the use of summary text mitigates the limitation, future studies could use a text encoder that can handle a long sequence to exploit the entire news article. Third, this study focused on developing a zero-shot classifier based on a CLIP-like dual encoder, which does not involve labeled data for training. The performance could be boosted by developing a fine-tuned classifier or few-shot prompt learning methods.

Ethics Statement

This study introduced CFT-CLIP, a contrastive learning framework for training an image-text multimodal encoder. For pretraining, this study used publicly available news articles shared by news media. While we tried to have a high-quality corpus for pretraining, it is possible that the model learned hidden biases in online news. Also, Since CFT-CLIP was updated from the pretrained CLIP weights, it may inherit the bias of CLIP. A user should be cautious about applying the method to problems in a general context and be aware of a potential bias. We have fewer privacy concerns because our study used openly accessible news data that may follow strict internal guidelines according to journalism principles. The NELA-GT-2021 was shared under the license of CC BY-NC 4.0, and the BBC corpus was shared under the MIT license. We will share NewsTT with CC BY NC 4.0. Some of the text was edited using AI assistants, such as ChatGPT and Grammarly.

Acknowledgements

Kunwoo Park is the corresponding author. This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (IITP-2024-RS-2022-00156360).

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14940–14949.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text retrieval: A survey on recent research and development. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22) Survey Track*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Hyewon Choi, Yejun Yoon, Seunghyun Yoon, and Kunwoo Park. 2022. How does fake news use a thumbnail? CLIP-based multimodal detection on the unrepresentative news image. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 86–94, Dublin, Ireland. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maurício Gruppi, Benjamin D Horne, and Sibel Adalı. 2022. Nela-gt-2021: A large multi-labelled news dataset for the study of misinformation in news articles. *ArXiv preprint*, abs/2203.05659.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. 2023. Mixgen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 379–389.

Alfred Hermida, Fred Fletcher, Darryl Korell, and Donna Logan. 2012. Share, like, recommend: Decoding the social media news consumer. *Journalism studies*, 13(5-6):815–824.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Hélène Joffe. 2008. The power of visual material: Persuasion, emotion and identification. *Diogenes*, 55(1):84–93.

Chei Sian Lee and Long Ma. 2012. News sharing in social media: The effect of gratifications and prior experience. *Computers in human behavior*, 28(2):331–339.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024. Gecko: Versatile text embeddings distilled from large language models.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022b. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Andreas Lommatzsch, Benjamin Kille, Özlem Özgöbek, Yuxiao Zhou, Jelena Tešić, Cláudio Bartolomeu, David Semedo, Lidia Pivovarova, Mingliang Liang, and Martha Larson. 2022. Newsimages: addressing the depiction gap with an online news dataset for text-image rematching. In *Proceedings of the 13th ACM Multimedia Systems Conference*, page 227–233, New York, NY, USA. Association for Computing Machinery.

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.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Factify: A multi-modal fact verification dataset. In *Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY)*.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 16–25.

Eryn J Newman, Maryanne Garry, Daniel M Bernstein, Justin Kantner, and D Stephen Lindsay. 2012. Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review*, 19:969–974.

Eryn J Newman and Lynn Zhang. 2020. How nonprobative photos shape belief. *Cognitive Science*, page 90.

Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. EASE: Entity-aware contrastive learning of sentence embedding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3870–3885, Seattle, United States. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv preprint*, abs/1807.03748.

Nelleke Oostdijk, Hans van Halteren, Erkan Başar, and Martha Larson. 2020. The connection between the text and images of news articles: New insights for multimedia analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4343–4351, Marseille, France. European Language Resources Association.

OpenAI. 2023. Gpt-4v(ision) system card.

Kunwoo Park, Haewoon Kwak, Jisun An, and Sanjay Chawla. 2021a. How-to present news on social media: A causal analysis of editing news headlines for boosting user engagement. In *ICWSM*, pages 491–502.

Kunwoo Park, Zhufeng Pan, and Jungseock Joo. 2021b. Who blames or endorses whom? entity-to-entity directed sentiment extraction in news text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4091–4102.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Anku Rani, S.M Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY-5WQA: 5W aspect-based fact verification through question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10421–10440, Toronto, Canada. Association for Computational Linguistics.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kiwon Seo. 2020. Meta-analysis on visual persuasion–does adding images to texts influence persuasion. *Athens Journal of Mass Media and Communications*, 6(3):177–190.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

The Associate Press. 2022. *The AP Stylebook: 2022-2024*. Basic Books.

Yash Verma, Anubhav Jangra, Raghvendra Verma, and Sriparna Saha. 2023. Large scale multi-lingual multi-modal summarization dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3620–3632, Dubrovnik, Croatia. Association for Computational Linguistics.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

Dong Yao, Zhou Zhao, Shengyu Zhang, Jieming Zhu, Yudong Zhu, Rui Zhang, and Xiuqiang He. 2022. Contrastive learning with positive-negative frame mask for music representation. In *Proceedings of the ACM Web Conference 2022*, pages 2906–2915.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. 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 2099–2108, Hong Kong, China. Association for Computational Linguistics.

Appendix

A Unlabeled data characteristics

Table A1 shows the data characteristics for the unlabeled train dataset. In reference to Table 1, we found that the labeled data follows a similar distribution to the unlabeled NELA across various dimensions. The labeled summary text tends to have different characteristics from the unlabeled BBC summary dataset because the former was generated by GPT yet the latter is written by the news editor.

	NELA	BBC	
	Title	Title	Summary
Words	14.3	9.3	25.4
Nouns	3.24	2.56	5.22
Verbs	1.67	1.07	2.44
Adjectives	0.94	0.61	1.54
Named entities	2.14	1.73	2.71

Table A1: Data distribution

B Configuration details

We ran experiments on a machine equipped with AMD Ryzen Threadripper Pro 5975WX CPU, three Nvidia RTX A6000 GPUs (48GB per GPU), and 256GB RAM. We trained the models with the AdamW optimizer with the initial learning rate of 1e-4, updated by the cosine annealing scheduler. The minibatch size is 128. The temperature τ in the loss equation is 0.05, following Gao et al. (2021). Other hyperparameters were optimized by random search using a validation set. Model training was early-stopped when the validation loss was not decreased five times consecutively, measured for every 20 iterations. The experiments were conducted on Python 3.9, Pytorch 1.10.1, Transformers 4.29.2, LAVIS 1.0.2, and SentenceTransformer 2.2.2. Five random seeds were used for repeated experiments: 0, 1, 2, 3, and 4. The temperature used for adjusting the masked token prediction is set as 2.0.

The pretrained model checkpoints used for the experiments are as follows:

- CLIP: `https://huggingface.co/openai/clip-vit-large-patch14`
- BLIP: blip_base (LAVIS)
- BLIP-2: blip2_pretrain (LAVIS)
- BLIP-2+SBERT: pretrain_opt2.7b (LAVIS), all-MiniLM-L6-v2 (SentenceTransformer)
- Named Entity Recognizer: `https://huggingface.co/spacy/en_core_web_trf`
- Masked Language Model: `https://huggingface.co/bert-base-uncased`

C Ablation experiments

Reference text We compared the performance of using news title and summary text as reference text T in the zero-shot classifier. Table A2 presents the results, showing that using the summary text

could achieve a better outcome. Thus, we used the summary text for inference.

T	F1	Spearman
Summary	**0.815**±0.003	**0.491**±0.005
Title	0.759±0.003	0.386±0.003

Table A2: Performance by the reference text type during inference

Transfer learning To understand whether the proposed contrastive learning framework can lead to a better model than the standard CLIP objective without transfer learning, we conducted experiments that update the parameters of a vision and language bi-encoder from scratch by the CLIP and CFT-CLIP objectives. We used the transformer models with the same architecture but with randomly initialized parameters. We used the BBC dataset with summary text for training, and all parameters were updated until the 20th epoch. The other settings remained the same as those used in the main experiment. Table A3 shows the results, indicating that CFT-CLIP outperformed CLIP with an f1 of 0.703 and a Spearman coefficient of 0.06. This suggests that contrasting with the counterfactual text can make the vision language bi-encoder learn the cross-modal matching ability for the target task. However, its performance was lower than that of the pretrained CLIP reported in Table 2, which achieved an f1 of 0.763 and a Spearman coefficient of 0.409. Given that the pretrained CLIP was trained on a web-scaled dataset, we guess the performance degradation originated from the scale of the pretraining corpus.

Model	F1	Spearman
CFT-CLIP	**0.703±0.008**	**0.060±0.016**
CLIP	0.625±0.005	0.045±0.007

Table A3: Results without transfer learning

D Data examples

D.1 Failed prediction

Table A1 presents two examples where the CFT-CLIP model made a failed prediction. The first example presents the false positive case where the model returns a high similarity score for the unrepresentative thumbnail. The example on the right shows an error case where referring to external knowledge is required. The person in the image is Glenn Youngkin, the governor of Virginia in the US. A model could not return a high similarity score without knowing his background.

D.2 Counterfactual text

Table A4 presents several examples derived from our counterfactual text generation method. Overall, the proposed method successfully generates the counterfactual text by selecting appropriate entities and substituting them with other entities. On the other hand, there are several cases where the generated token breaks the grammar. For instance, in the fourth example, the term 'Oisin Murphy' was replaced by 'offensive.' Such anomalies may arise from the sampling process employed, which aims to prevent the recreation of the original text. Despite its imperfect structure, the generated sentence can still serve as a hard negative sample during contrastive update, given that its general context is preserved.

E API usage details

We used OpenAI API to use GPT 3.5-Turbo. We obtained news summaries for the labeled dataset and generated counterfactual text for the summary text of the BBC unlabeled pretraining corpus. In total, the API call cost $13.33. Below are the prompts used in the experiments.

News summarization

> Article:{text}
>
> Summarize the article in one sentence.

Counterfactual text generation

> Create a counterfactual news summary by modifying the actors of news events:{text}
>
> Answer in JSON. The JSON should be a string of dictionaries whose keys are "counterfactual".

F Annotation details and guidelines

We hired two male and one female student from Soongsil University. The annotators were trained by using the guidelines in Table A5. The original guideline was in another language, and we present its English-translated version. All the annotators were paid $0.1 per example.

Figure A1: More error examples

Original text	Generated text
Op-Ed: Memo to Saddleback Church: Replacing Pastor Rick Warren is a minefield	Op-Ed: Memo to Saddleback Church: Replacing Pastor Parker is a minefield
'The greatest striker': Gerd Müller, legendary German forward, dies aged 75	'The greatest striker': Joseph, legendary German forward, dies aged 75
Matt Gaetz and wingman facing 'mutually assured destruction' after confession letter: legal expert	Pennant and wingman facing 'mutually assured destruction' after confession letter: legal expert
William Buick treble sets up Flat jockeys' title race for dramatic finish as gap closes on Oisin Murphy	Davidson treble sets up Flat jockeys' title race for dramatic finish as gap closes on offensive
Michael Douglas says it was 'uncomfortable' for him and Catherine Zeta-Jones to share Mallorcan home with his ex	Novella says it was 'uncomfortable' for him and Catherine Zeta-Jones to share Mallorcan home with his ex
Experts say Jussie Smollett is in 'matrix of arrogance' as he awaits sentencing	Experts say Toni is in 'matrix of arrogance' as he awaits sentencing
Ted Nugent tests positive for coronavirus after calling pandemic a 'scam'	Danny tests positive for coronavirus after calling pandemic a 'scam'
Jennifer Aniston Explained How Therapy Helps Her Deal With The "Tough Stuff" Of Being Famous	Charles Explained How Therapy Helps Her Deal With The "Tough Stuff" Of Being Famous
Thousands mark anniversary of Kremlin critic Nemtsov's murder	Thousands mark anniversary of Kremlin critic pastor's murder
Nets Disregard AG Garland Grilled in Hearing for Targeting Parents	Nets Disregard AG Cicero Grilled in Hearing for Targeting Parents
Travis Kelce Is Borderline Unrecognizable Without Facial Hair	Annie Is Borderline Unrecognizable Without Facial Hair
Billionaire Ken Griffin bought a copy of the US Constitution for $43.2m because his son asked him to	Billionaire Dublin bought a copy of the US Constitution for $43.2m because his son asked him to
Desperate Chuck Todd Hopes Trump Will Deflect Media 'Spotlight' From Dem 'Problems'	Desperate Joe Hopes Radha Will Deflect Media 'Spotlight' From Dem 'Problems'
Chris Cuomo's Book Contract Dropped By HarperCollins	Harriet Book Contract Dropped By HarperCollins
Jen Psaki shoots down a reporter comparing Biden to his predecessor: Trump suggested 'people inject bleach'	person shoots down a reporter comparing virus to his predecessor: Walden suggested 'people inject bleach'
What's the Deal With Gavin Newsom? 5 Plausible Theories To Explain His Mysterious Hiatus	What's the Deal With diplomacy? 5 Plausible Theories To Explain His Mysterious Hiatus

Table A4: Counterfactual text generation examples

Task overview:
In this task, you are asked to answer whether a given news image represents the actors of the news events.

Instruction:
- Q1. Identify news actors in the text. The actors can be expressed as named entities, proper nouns, or common nouns.
- Q2. Does the image display the news actors identified in Q1?
- Q3. Identify the visually presented news actors in the image.

Examples:

Title: Trump Announces New Impeachment Defense Team
Summary: Former President Donald Trump announces a new legal defense team, led by attorneys David Schoen and Bruce Castor Jr., for his upcoming second impeachment trial following the disbanding of the original team due to disagreements over legal strategy, with the new team focusing on representing Trump and the United States Constitution without commenting on election fraud allegations.

- Q1: Trump, Impeachment Defense Team, David Schoen, Bruce Castor Jr., impeachment trial, disagreements, legal strategy, United States Constitution, election fraud allegations
- Q2: Y
- Q3: Trump

(More examples)

Table A5: Translated annotation guideline