# Identification of Necessary Semantic Undertakers in the Causal View for Image-Text Matching

**Huatian Zhang, Lei Zhang, Kun Zhang, Zhendong Mao***

University of Science and Technology of China, Hefei, China
{huatianzhang, kkzhang}@mail.ustc.edu.cn, {leizh23, zdmao}@ustc.edu.cn

## Abstract

Image-text matching bridges vision and language, which is a fundamental task in multimodal intelligence. Its key challenge lies in how to capture visual-semantic relevance. Fine-grained semantic interactions come from fragment alignments between image regions and text words. However, not all fragments contribute to image-text relevance, and many existing methods are devoted to mining the vital ones to measure the relevance accurately. How well image and text relate depends on the degree of semantic sharing between them. Treating the degree as an effect and fragments as its possible causes, we define those indispensable causes for the generation of the degree as necessary undertakers, *i.e.*, if any of them did not occur, the relevance would be no longer valid. In this paper, we revisit image-text matching in the causal view and uncover inherent causal properties of relevance generation. Then we propose a novel theoretical prototype for estimating the probability-of-necessity of fragments, $\mathrm{PN}_f$, for the degree of semantic sharing by means of causal inference, and further design a Necessary Undertaker Identification Framework (NUIF) for image-text matching, which explicitly formalizes the fragment's contribution to image-text relevance by modeling $\mathrm{PN}_f$ in two ways. Extensive experiments show that our method achieves state-of-the-art on benchmarks Flickr30K and MSCOCO.

## 1 Introduction

Image-text matching aims to search for semantically relevant images given text or retrieve descriptive texts given image, which is a fundamental task in multimodal intelligence facilitating many applications, such as information database search and e-commerce recommendation. Despite considerable development in recent years, image-text matching remains the challenge in capturing visual-semantic relevance.

Extensive research has been done to study the semantic relevance from interactions of cross-modal contents. A common framework is aligning constituent fragments (image regions or text words) semantically and aggregating the resulted alignments accordingly. (Lee et al. 2018) proposed a cross-attention mechanism to capture all latent alignments by attending to regions and words with each other as context, and inspires a bunch of studies. (Zhang et al. 2020b; Wehrmann, Kolling, and Barros 2020; Chen and Luo 2020; Liu et al. 2020) constructed thoughtful aligning rules to capture fine interactions. (Diao et al. 2021) explored self-attention reasoning as an aggregation mechanism to enhance meaningful alignments. (Zhang et al. 2022a) assigned high confidence to image regions consistent with the global semantics in aggregating. (Pan, Wu, and Zhang 2023) proposed to eliminate redundant or irrelevant fragment alignments from the perspective of information coding. In general, not all fragments contribute to image-text relevance, and a large branch of existing methods is devoted to mining the vital ones to measure the relevance accurately.

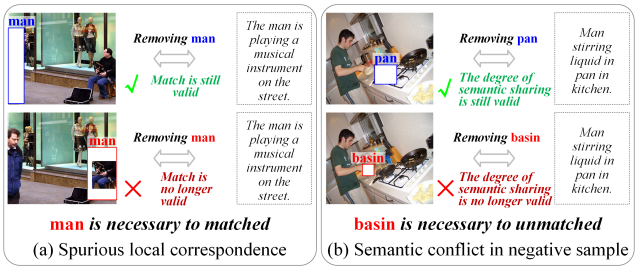



Figure 1: Illustration of necessary undertakers. (a) Although the man in blue box spuriously corresponds to text in fragmental aligning, removing it will not break image-text match. If the man in red box is removed, the matched relationship will no longer hold, *i.e.*, the man in red box is necessary. (b) Basin is the critical conflict that causes image-text unmatch. Removing the basin will affect the degree of semantic sharing between image and text, but the pan will not, *i.e.*, the basin is necessary to this unmatched relationship.

Normally, how well image and text relate depends on the degree to which they overlap into shared semantics. Fragments that contribute to image-text relevance are those that, if any of them did not occur, the relevance would be no longer valid, *i.e.*, they are necessary to image-text relevance. In other words, these fragments are *necessary undertakers* of the degree of semantic sharing between image and text. Although the unnecessary ones may also locally correspond to the other modality, identifying necessity can filter out such spurious correspondences from image-text relevance measure by their low necessity, to reduce the impact on match-

*Corresponding author.

ing. As shown in Fig. 1(a), even if the unnecessary man in blue is locally related to the text, the necessity suppresses its contribution to overall relevance. Meanwhile, as the redundancy of unnecessary fragments, excluding them from image or text can help to establish alignments more discriminatively, without altering the inherently shared semantics. Particularly for hard negatives, identifying necessity helps to focus on the evidentiary conflicts. As shown in Fig. 1(b), necessary basin region points out the semantic contradiction.

Treating the degree of semantic sharing as an effect and fragments as its possible causes, the necessary undertakers refer to those indispensable causes for the generation of the degree. Thereby, identifying necessary undertaker is equivalent to determining the probability that the fragment is the degree's cause. For this purpose, we aim to answer "What would happen to image-text relevance if a fragment did not occur?". From the perspective of causal inference (Pearl 2009; Glymour, Pearl, and Jewell 2016), the question hypothesizes an absence of fragment, and introduces a comparison on the degree of semantic sharing between actuality and the imaginary scenario where the fragment absents. To capture how the degree varies, we express semantic change of image or text caused by the absence of fragment by the fragment's semantic dependency, which collects regions or words that have direct semantic causalities with it. Then we structurally model the generation of image-text relevance to specify the functional relationships that connect semantic changes and the relevance, and further uncover two causal properties of relevance generation in matching, the exogeneity of semantic dependency and the matching monotonicity.

Based on the insights above, we propose a novel theoretical prototype for estimating the probability-of-necessity of fragments by means of causal inference, and further design a Necessary Undertaker Identification Framework (NUIF) for image-text matching, which quantitatively identifies necessary undertakers of the degree of semantic sharing between image and text in measuring overall relevance. Specifically, we first relate fragments between modalities to obtain vision-language alignments. Then we represent image and text adaptively by highlighting regions or words that are most semantically consistent with each other to which they aligned, to enable relevance measurement sensitive to such match-critical fragments so that it can clearly reflect the variation in how image and text overlap when semantics change. Finally, we quantify the probability-of-necessity of fragments counterfactually by relative variation in semantic overlapping after removing their semantic dependencies, and then aggregate the alignments queried by necessary fragments into image-text relevance for matching. Our framework explicitly formalizes the fragment's contribution to image-text relevance from a causal perspective, which achieves the goal more intuitively.

Our contributions are summarized as follows: (1) We revisit image-text matching in the causal view and, to the best of our knowledge, we are the first to propose a novel theoretical prototype for estimating the probability-of-necessity of fragments to undertake the degree of semantic sharing between image and text. (2) We propose a Necessary Undertaker Identification Framework (NUIF) that adaptively highlights match-critical fragments in representing image and text, and quantifies the probability-of-necessity of fragments counterfactually by relative variation in how image and text overlap after removing fragments' semantic dependencies. (3) The experimental results validate the effectiveness of our proposed method, and demonstrate that NUIF achieves state-of-the-art on benchmarks Flickr30K and MSCOCO.

## 2 Related Work

**Image-Text Matching.** To capture image-text relevance for matching from fine-grained interactions on fragments, extensive works have been proposed. Different from the research line that focuses on representing the holistic image or text to perform coarse cross-modal interaction (Chen et al. 2021; Yan, Yu, and Xie 2021; Li et al. 2022b; Fu et al. 2023), the research line examining fine-grained interactions attracts a lot of attention. One of the representative (Lee et al. 2018) proposed the cross-attention mechanism that aims to discover all region-word fragmental alignments and inspires a series of works (Wehrmann, Kolling, and Barros 2020; Chen and Luo 2020; Liu et al. 2020; Ji, Chen, and Wang 2021; Zhang et al. 2023a). Some works focused on exploiting more information, such as scene graph (Wang et al. 2020b), consensus knowledge (Wang et al. 2020a), and external pretrained knowledge (Wei et al. 2020; Qu et al. 2021; Yao et al. 2021), etc., to enhance cross-modal alignments. Another line of methods focused on constructing thoughtful aggregating rules to capture vital fragmental interactions. (Liu et al. 2020) and (Diao et al. 2021) explored the structure aligning between regions and words via graph neural network. (Zhang et al. 2022a) assigned confidence to regions to emphasize alignments queried by reliable ones in semantic relevance aggregation. (Zhang et al. 2022b) proposed the negative-aware attention to use the misaligned fragments explicitly. (Kim, Kim, and Kwak 2023) coded samples into a set of different embeddings that captures diverse semantics to handle ambiguity. (Pan, Wu, and Zhang 2023) proposed eliminating irrelevant alignments through cross-modal hard aligning based on coding theory.

**Causality in Computer Vision.** Causal inference (Pearl 2009; Glymour, Pearl, and Jewell 2016) has been widely applied to computer vision to gain insight into the intrinsic causal mechanism of tasks, including visual recognition (Wang et al. 2020c; Liu et al. 2022; Mao et al. 2022), semantic segmentation (Zhang et al. 2020a), scene graph generation (Tang et al. 2020), video analysis (Li et al. 2021; Liu et al. 2021), domain generalization (Lv et al. 2022; Chen et al. 2023a), object navigation (Zhang et al. 2023b), etc. In multimodal machine learning, (Yang et al. 2021) alleviated the dataset bias in image captioning based on the backdoor and frontdoor adjustment principles. (Wei et al. 2022) synthesized counterfactual samples to augment training data for image-text matching. (Chen et al. 2023b) proposed a counterfactual samples synthesizing and training strategy to improve visual-explainable and question-sensitive abilities of visual question answering. (Zang et al. 2023) captured video features causally related to question to restrain redundant language semantics on question answering. In this paper, we examine image-text matching in the causal view, to estimate

the probability-of-necessity of fragments to undertake the degree of semantic sharing between image and text, in order to identify the necessary undertakers quantitatively in measuring image-text relevance.

## 3 Image-Text Matching in the Causal View

We start by structurally modeling the generation of image-text relevance from the perspective of causality, to understand the semantic change in image or text when a fragment absents, and extract the causal properties inherent in the relevance generation. Then we derive a theoretical prototype for estimating the probability-of-necessity of fragments to the degree of semantic sharing by counterfactual means.

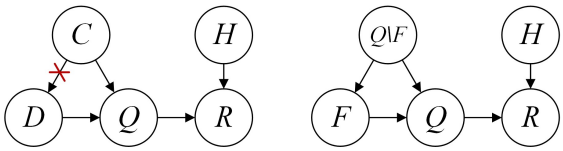
### 3.1 Structual Modeling

Given an image or text, image-text matching is to rank candidates (texts or images) based on semantic relevance between modalities. Generally, a limited number of visual concepts that have semantic dependencies with a region can host almost losslessly visual context related to this region in the image. The rest have no direct cause-and-effect on whether the region occurs or not. In the text, the words in a syntactic component or phrase are often linguistically interdependent and work together as a fine-grained semantic unit. Moreover, specific meanings activated for the region or word are constrained by the visual or linguistic context that it is involved. These facts inspire us to partition an image or a text for each constituent fragment into semantic dependency, which includes the fragment itself and gathers up regions or words that have direct semantic causalities with the fragment, and semantic complement. When a fragment is removed, the semantics in the image or text that emerge from its dependency will no longer hold due to causal disruption.

As shown in Fig. 2(a), we build a causal graph to formalize the causalities among variables: image or text query $Q$, the semantic dependency $D$ and complement $C$ of a fragment $F$, heteromodal candidate $H$, and image-text relevance $R$ between $Q$ and $H$, each vertex corresponds to a variable, each edge denotes the cause-and-effect relationship between its end-vertices. Concretely, $Q \rightarrow R \leftarrow H$ denotes that the relevance $R$ is determined by how well the semantics of query $Q$ and candidate $H$ overlap. $D \rightarrow Q \leftarrow C$ indicates that, from the perspective of fragment $F$, image or text query $Q$ is composed of its dependency $D$ and complement $C$ organically. $D$ gathers up the fragments that have direct semantic causalities with $F$, that is, it recapitulates the context of $F$ in $Q$. $D \perp\!\!\!\perp C$ means that concepts in $C$ cannot cause the occurrence or not of $D$ in terms of semantic logic. While, as shown in Fig. 2(b), $F$ alone cannot be free from the calling-up from some other fragments in $Q\backslash F$.

### 3.2 Necessity Estimation

In causal theory, given an event $Y$ and its possible cause $X$, a counterfactual interpretation of causation that effect $Y$ would not have occurred in the absence of $X$ captures how necessary the cause $X$ is for the production of $Y$, *i.e.*, probability-of-necessity. The potential response of $Y$ to hypothetical action $X = x$ is denoted as $Y_x$, and $y_x$ indicates




(a) Semantic dependency (b) Fragmented view

Figure 2: The causal graphs of image-text matching. (a) There is no cause-and-effect between $D$ and the remaining complement $C$ from $F$'s point of view. (b) Fragment $F$ may semantically depend on its context fragments in $Q\backslash F$.

that $Y$ would be $y$ if $X$ had been $x$. Then, under binary logic, the probability-of-necessity can be defined counterfactually as $\mathrm{PN} := P\left(y'_{x'} \mid x, y\right)$, standing for the probability of $y'_{x'}$ given that $x$ and $y$ did occur, where $x$ and $y$ denote respectively the events $X$ = true and $Y$ = true, otherwise false. Under certain assumptions, the quantity of probability-of-necessity can be estimated from observational data facts.

To estimate the probability-of-necessity of fragments for the degree of semantic sharing between image and text, we first uncover two inherent causal properties in the generation of image-text relevance as follows.

**Exogeneity of Semantic Dependency.** In the causal graph, if variable $D$ is fixed to $d$, the variation in potential response of $R$ to $D = d$, $R_d$, will be dominated by other variables that can affect $R$. However, $D$ and $R$ have no common ancestor variable, *i.e.*, no confounding. Hence variables capable of transmitting variations to $R$ are independent of $D$, and so is $R_d$. For the semantic dependency of a fragment $f$, the way $R$ would respond to its occurrence $D = o_d$ or absence $D = o'_d$ is independent of the actual value of $D$, thus:

$$\left\{R_{o_d}, R_{o'_d}\right\} \perp\!\!\!\perp D. \tag{1}$$

In causal terms, the dependency $D$ is exogenous relative to image-text relevance $R$.

**Matching Monotonicity.** For images, peeling away neither salient regions together with their dependent context nor trivial backgrounds will render the image that does not match a description match better. Similarly, in texts, masking out syntactic components will reduce the descriptive semantics and make the text sketchy or even blur its logic, thus the masked text ought not to be more relevant to image candidates. It can be summed up as the absence of semantic dependency, $D = o'_d$, cannot make query $Q$ that does not match heteromodal candidate $H$ turn to match. Furthermore, let $M$ denotes matching degree between $Q$ and $H$, in the binary case, $m$ for $M$ = true and $m'$ the opposite. Then:

$$m_{o'_d} \wedge m'_{o_d} = \text{false}, \tag{2}$$

that is, the matching degree $M$ is monotonic relative to the occurrence of semantic dependency $D$.

Putting the insights above together, the probability-of-necessity can be quantified for identifying necessary undertakers in image-text matching.

**Theorem 1** (Necessary Undertaker Identification)**.** *In image-text matching, for a fragment $f$ in image or text, with*

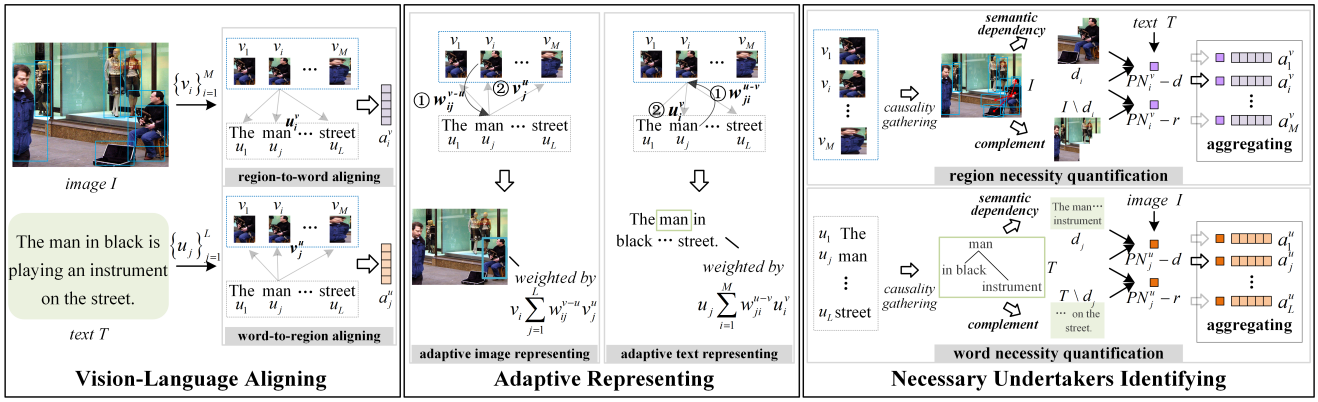



Figure 3: Illustration of our proposed NUIF. The framework consists of three modules: vision-language aligning, adaptive representing, and necessary undertakers identifying. The symbols ① and ② in adaptive representing indicate the order of computation. We model the probability-of-necessity, $\mathrm{PN}_f$, for the degree of semantics sharing in two ways, $\mathrm{PN}_f$-$d$ and $\mathrm{PN}_f$-$r$.

*semantic dependency $d$, its probability-of-necessity of undertaking the degree of semantics sharing between image and text, $\mathrm{PN}_f$, can be quantified by $(P(m \mid o_d) - P(m \mid o'_d))/P(m \mid o_d)$, where $m$ indicates the event that image and text match, $o_d$ denotes $f$'s semantic dependency occurs and $o'_d$ for its absence.*

*Proof.* See Appendix A for details. □

Note that it does not need to guarantee that $M$ is binary in Thm. 1. For the case of continuous $M$, Thm. 1 can also measure the relative weakening in relevance between image-text pairs with $o_d$ and $o'_d$ as the probability-of-necessity.

## 4 The Proposed Implementation

We then elaborate on the implementation of $\mathrm{PN}_f$ in Thm.1. Specifically, as shown in Fig. 3, given image $I = \{\boldsymbol{v}_i | i = 1, 2, \ldots, M, \boldsymbol{v}_i \in \mathbb{R}^D\}$, where $\boldsymbol{v}_i$ denotes detected salient region, and text $T = \{\boldsymbol{u}_j | j = 1, 2, \ldots, L,\ \boldsymbol{u}_j \in \mathbb{R}^D\}$, where $\boldsymbol{u}_j$ is word embedding, we design a Necessary Undertaker Identification Framework as: (1) Vision-Language Aligning: We relate regions and words to obtain visual-semantic alignments; (2) Adaptive Representing: We adaptively represent the image and text in matching by emphasizing regions or words that are semantically bijective with what they aligned in another modality, to enable relevance measure sensitive to such match-critical fragments so that the measure can acutely reflect the change in image-text semantic overlap; (3) Necessary Undertakers Identifying: We model the $\mathrm{PN}_f$ in two ways, one to measure the difference $\Delta P = P(m \mid o_d) - P(m \mid o'_d)$ integrally and model the $\mathrm{PN}_f$ as $\Delta P / P(m \mid o_d)$, named $\mathrm{PN}_f$-$d$, and the other to measure the ratio $P^{\mathrm{r}} = P(m \mid o'_d)/P(m \mid o_d)$ as a whole and model the $\mathrm{PN}_f$ as $1 - P^{\mathrm{r}}$, named $\mathrm{PN}_f$-$r$.

### 4.1 Vision-Language Aligning

To capture visual-semantic relevance at the fragment level, we obtain the semantically related fragments for one in another modality through the cross-attention mechanism. For region $\boldsymbol{v}_i$, we measure its attention weight on word $\boldsymbol{u}_j$ by $w_{ij}^{\mathrm{v}} = \frac{e^{(\lambda \hat{c}_{ij})}}{\sum_{j=1}^{N} e^{(\lambda \hat{c}_{ij})}}$, $\hat{c}_{ij} = [c_{ij}]_+ / \sqrt{\sum_{i=1}^{M} [c_{ij}]_+^2}$, where $\lambda$ is a constant temperature parameter and $c_{ij}$ is the cosine similarity between region $\boldsymbol{v}_i$ and word $\boldsymbol{u}_j$, and aggregate $\boldsymbol{v}_i$'s relevant words as its linguistic context $\boldsymbol{u}_i^{\mathrm{v}} = \sum_{j=1}^{L} w_{ij}^{\mathrm{v}} \boldsymbol{u}_j$. Then we embody vision-language alignment as a vector-valued distance between $\boldsymbol{v}_i$ and its attended context $\boldsymbol{u}_i^{\mathrm{v}}$ as:

$$\boldsymbol{a}_i^{\mathrm{v}} = l_2\text{-normalized}(\tanh(W_{\mathrm{v}} |\boldsymbol{v}_i - \boldsymbol{u}_i^{\mathrm{v}}|^2)), \tag{3}$$

where $W_{\mathrm{v}} \in \mathbb{R}^{P \times D}$ denotes a learnable projection matrix. It can be said that alignment $\boldsymbol{a}_i^{\mathrm{v}}$ is queried by $\boldsymbol{v}_i$. Similarly, the alignment $\boldsymbol{a}_j^{\mathrm{u}}$ queried by word $\boldsymbol{u}_j$ is $\boldsymbol{a}_j^{\mathrm{u}} = l_2\text{-normalized}(\tanh(W_{\mathrm{u}} |\boldsymbol{u}_j - \boldsymbol{v}_j^{\mathrm{u}}|^2))$, where $\boldsymbol{v}_j^{\mathrm{u}}$ is the visual context of $\boldsymbol{u}_j$and aggregated from regions with $\sum_{i=1}^{M} w_{ji}^{\mathrm{u}} \boldsymbol{v}_i$.

### 4.2 Adaptive Representing

To make relevance measurement more responsively reflect the variation in how image and text overlap as their semantics change, *e.g.*, from $o_d$ to $o'_d$, we adaptively represent the image and text by highlighting regions or words that are most semantically consistent with each other, *i.e.*, bijective with, to which they aligned in another modality. Such regions or words are match-critical since they are exactly the grounding bases of image-text semantic overlapping.

In detail, to represent the image $I$ in matching $I$ and text $T$, for region $\boldsymbol{v}_i$, we first obtain its most semantically aligned word as $\sum_{j=1}^{L} w_{ij}^{\mathrm{v-u}} \boldsymbol{u}_j$ by $\boldsymbol{w}_i^{\mathrm{v-u}} = [w_{i1}^{\mathrm{v-u}}, w_{i2}^{\mathrm{v-u}}, \ldots, w_{iL}^{\mathrm{v-u}}] = \mathrm{softmax}(\tau \cdot \log(\boldsymbol{w}_i^{\mathrm{v}}))$, where $\boldsymbol{w}_i^{\mathrm{v}}$ is the attention distribution of $\boldsymbol{v}_i$ on text words in Sec. 4.1, and $\tau$ is a temperature parameter. Then, we measure the likelihood that $\boldsymbol{v}_i$'s linguistic counterpart exists in the text $T$ to indicate the degree to which the $\boldsymbol{v}_i$ is match-critical by:

$$s_i^{\mathrm{v}} = \boldsymbol{v}_i \cdot \sum_{j=1}^{L} w_{ij}^{\mathrm{v-u}} \boldsymbol{v}_j^{\mathrm{u}}, \tag{4}$$

which is the similarity between $\boldsymbol{v}_i$ and $\sum_{j=1}^{L} w_{ij}^{\mathrm{v-u}} \boldsymbol{v}_j^{\mathrm{u}}$ that serves as the visual context of $\boldsymbol{v}_i$'s semantically aligned

word $\sum_{j=1}^{L} w_{ij}^{\mathrm{v-u}} \boldsymbol{u}_j$. It measures the similarity between $\boldsymbol{v}_i$ and the regions aligned by word $\sum_{j=1}^{L} w_{ij}^{\mathrm{v-u}} \boldsymbol{u}_j$. The more similar the two are, the more likely there is a linguistic counterpart of $\boldsymbol{v}_i$ in the text $T$, *i.e.*, the more match-critical $\boldsymbol{v}_i$ is. For the image $I$, we obtain $\boldsymbol{s}^{\mathrm{v}} = [s_1^{\mathrm{v}}, s_2^{\mathrm{v}}, \ldots, s_M^{\mathrm{v}}]$ and represent image $I$ by:

$$\boldsymbol{i} = \sum_{i=1}^{M} \mathrm{softmax}(\boldsymbol{s}^{\mathrm{v}})_i \boldsymbol{v}_i, \tag{5}$$

which highlights match-critical regions within $I$ through weights $\boldsymbol{s}^{\mathrm{v}}$ adaptively. Likewise, the text $T$ is represented as $\boldsymbol{t} = \sum_{j=1}^{L} \mathrm{softmax}(\boldsymbol{s}^{\mathrm{u}})_j \boldsymbol{u}_j$, where $s_j^{\mathrm{u}}$ is similar to Eq. 4.

### 4.3 Necessary Undertakers Identifying

We first gather the semantic dependency of individual fragments as follows. In image $I$, the regions that are semantically dependent on region $\boldsymbol{v}_i$ will tend to interact with it, which usually manifests as spatial proximity. For $\boldsymbol{v}_i$, we regard regions that are relatively close to $\boldsymbol{v}_i$, together with $\boldsymbol{v}_i$ itself as its semantic dependency $d_i$. In text $T$, the phrase to which $\boldsymbol{u}_j$ belongs is naturally its dependency $d_j$, while the word not belonging to any phrase is its own dependency.

Here it is ready for identifying the necessary undertakers. We model the $\mathrm{PN}_f = (P(m \mid o_d) - P(m \mid o'_d))/P(m \mid o_d)$ in two ways as follows.

**PN$_f$-d.** The **d**ifference $\Delta P = P(m \mid o_d) - P(m \mid o'_d)$ expresses the variance in how well image and text match caused by removing $f$'s semantic dependency, and can be measured integrally as the relevance between the content emerged from dependency and image or text candidate instance. For region $\boldsymbol{v}_i$ with semantic dependency $d_i$, we represent the dependency as $\boldsymbol{d}_i = \sum_{k \in \mathrm{idx}_i} \mathrm{softmax}(\boldsymbol{s}^{\mathrm{v}}_{\mathrm{idx}_i})_k \boldsymbol{v}_k$, where $\mathrm{idx}_i$ denotes the index set of $d_i$. Then we measure the $\Delta P$ w.r.t. $\boldsymbol{v}_i$, $\Delta P_i^{\mathrm{v}}$, as:

$$\Delta P_i^{\mathrm{v}} = (1 + \frac{1}{N} \sum_{n=1}^{N} \frac{(\boldsymbol{d}_i)_n \cdot \boldsymbol{t}_n}{\|(\boldsymbol{d}_i)_n\|_2 \, \|\boldsymbol{t}_n\|_2})/2, \tag{6}$$

which is block-wise cosine similarity with $N$ blocks. It reduces noise in similarity measure through the refinement on dimensions. Then we measure $P(m \mid o_{d_i})$, the probability that image and text match, by similarity between $\boldsymbol{i}$ and $\boldsymbol{t}$:

$$P(m \mid o_{d_i}) = (1 + \frac{1}{N} \sum_{n=1}^{N} \frac{\boldsymbol{i}_n \cdot \boldsymbol{t}_n}{\|\boldsymbol{i}_n\|_2 \, \|\boldsymbol{t}_n\|_2})/2. \tag{7}$$

Further, we model the probability-of-necessity of region $\boldsymbol{v}_i$ as $\mathrm{PN}_i^{\mathrm{v}}$-$d = \Delta P_i^{\mathrm{v}} / P(m \mid o_{d_i})$.

Similarly, we measure the $\mathrm{PN}_j^{\mathrm{u}}$-$d$ of the word $\boldsymbol{u}_j$ with semantic dependency $d_j$ as $\Delta P_j^{\mathrm{u}} / P\left(m \mid o_{d_j}\right)$, where $\Delta P_j^{\mathrm{u}} = \max_{k \in \mathrm{idx}_j} ((1 + \frac{1}{N} \sum_{n=1}^{N} \frac{(\boldsymbol{u}_k)_n \cdot \boldsymbol{t}_n}{\|(\boldsymbol{u}_k)_n\|_2 \|\boldsymbol{t}_n\|_2})/2)$, which is the largest relevance variation can be caused by removing word in $d_j$.

**PN$_f$-r.** The **r**atio $P(m \mid o'_d)/P(m \mid o_d)$ means the retention rate of the probability that image and text match when the semantic dependency $d$ changes from occurrence $o_d$ to absence $o'_d$. It implies the level of likeness between image-text shared semantics under $o_d$ and $o'_d$. For region $\boldsymbol{v}_i$, we represent its semantic complement in image $I$ as $\boldsymbol{c}_i = \sum_{k \in \mathrm{idx}_i^{\mathrm{c}}} \mathrm{softmax}(\boldsymbol{s}^{\mathrm{v}}_{\mathrm{idx}_i^{\mathrm{c}}})_k \boldsymbol{v}_k$, where $\mathrm{idx}_i^{\mathrm{c}}$ denotes the complement of $\mathrm{idx}_i$ in image $I$. Then we formulate the $P\left(m \mid o'_{d_i}\right)$ through vision-language aligning in Eq.3 between $\boldsymbol{c}_i$ and $\boldsymbol{t}$:

$$\begin{aligned} P\left(m \mid o'_{d_i}\right) &= \boldsymbol{a}^{\mathrm{v}}_{I \setminus d_i} \\ &= l_2\text{-normalized}(\tanh(W_{\mathrm{v}} |\boldsymbol{c}_i - \boldsymbol{t}|^2)), \end{aligned} \tag{8}$$

and the $P(m \mid o_{d_i})$ through the aligning between $\boldsymbol{i}$ and $\boldsymbol{t}$:

$$\begin{aligned} P(m \mid o_{d_i}) &= \boldsymbol{a}^{\mathrm{v}}_{I} \\ &= l_2\text{-normalized}(\tanh(W_{\mathrm{v}} |\boldsymbol{i} - \boldsymbol{t}|^2)). \end{aligned} \tag{9}$$

Further, we measure the ratio $P(m \mid o'_d)/P(m \mid o_d)$ by:

$$P_i^{\mathrm{r}} = (1 + \boldsymbol{a}^{\mathrm{v}}_{I \setminus d_i} \cdot \boldsymbol{a}^{\mathrm{v}}_{I})/2, \tag{10}$$

which means the projection of $\boldsymbol{a}^{\mathrm{v}}_{I \setminus d_i}$ onto $\boldsymbol{a}^{\mathrm{v}}_{I}$. Then we model the probability-of-necessity of $\boldsymbol{v}_i$ as $\mathrm{PN}_i^{\mathrm{v}}$-$r = 1 - P_i^{\mathrm{r}}$.

Likewise, we measure the $\mathrm{PN}_j^{\mathrm{u}}$-$r$ of the word $u_j$ with semantic dependency $d_j$ as $1 - P_j^{\mathrm{r}}$, where $P_j^{\mathrm{r}} = (1 + \boldsymbol{a}^{\mathrm{u}}_{T \setminus d_j} \cdot \boldsymbol{a}^{\mathrm{u}}_{T})/2$, similar to Eq. 10.

Then we aggregate the vision-language alignments queried by the necessary regions through $\mathrm{PN}^{\mathrm{v}}$-$d$ or $\mathrm{PN}^{\mathrm{v}}$-$r$ as $\boldsymbol{a}_I = \sum_i^M \mathrm{softmax}(\mathrm{PN}^{\mathrm{v}})_i \boldsymbol{a}_i^{\mathrm{v}}$, and the alignments queried by the necessary words through $\mathrm{PN}^{\mathrm{u}}$-$d$ or $\mathrm{PN}^{\mathrm{u}}$-$r$ as $\boldsymbol{a}_T = \sum_j^L \mathrm{softmax}(\mathrm{PN}^{\mathrm{u}})_j \boldsymbol{a}_j^{\mathrm{u}}$, then incorporate them into image-text relevance $r$ by:

$$r(I, T) = \tanh(\boldsymbol{w}_r([\boldsymbol{a}_I : \boldsymbol{a}_T])), \tag{11}$$

where $\boldsymbol{w}_r \in \mathbb{R}^{1 \times 2P}$ is a learnable vector, and the $[:]$ denotes the concatenation operation.

### 4.4 Training

**Feature Encoder.** For a fair comparison with previous works, we use the ROI features of pre-trained object detector as detected regions, and transform them to $D$-dimensional $\boldsymbol{v}_i$ via linear projection. For texts, we employ two types of extractors, Bi-GRU and pre-trained BERT (Kenton and Toutanova 2019). When using Bi-GRU, the embedding of the $i$-th word, $\boldsymbol{u}_j$, is averaged from its forward and backward hidden states. When using BERT, we linearly map its output hidden states to $D$-dimensional embeddings.

**Objective Function.** Ranking objectives are adopted in image-text matching widely to force matched image-text pairs close to each other and pull unmatched ones away. We use the bi-directional triplet loss, focusing on the hardest negatives in-batch for efficiency:

$$\begin{aligned} \mathcal{L}(I, T) = &\left[\alpha - r(I, T) + r\left(I, T_h^-\right)\right]_+ \\ &+ \left[\alpha - r(I, T) + r\left(I_h^-, T\right)\right]_+, \end{aligned} \tag{12}$$

where $\alpha$ is a margin constraint, $[\cdot]_+ = \max(x, 0)$. $I_h^- = \mathrm{argmax}_{I^- \neq I} r(I^-, T)$, and $T_h^- = \mathrm{argmax}_{T^- \neq T} r(I, T^-)$ are the hardest negatives, given positive matched $I$ and $T$.

## 5 Experiments

### 5.1 Datasets and Evaluation Metrics

We evaluate the proposed framework on Flickr30K (Young et al. 2014) and MSCOCO (Lin et al. 2014) datasets.

| Methods | Flickr30K | | | | | | | MSCOCO 1K | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IMG → TXT | | | TXT → IMG | | | rSum | IMG → TXT | | | TXT → IMG | | | rSum |
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| **BUTD + Bi-GRU** | | | | | | | | | | | | | | |
| GSMN(Liu et al. 2020) | 76.4 | 94.3 | 97.3 | 57.4 | 82.3 | 89.0 | 496.8 | 78.4 | 96.4 | 98.6 | 63.3 | 90.1 | 95.7 | 522.5 |
| GPO*(Chen et al. 2021) | 76.5 | 94.2 | 97.7 | 56.4 | 83.4 | 89.9 | 498.1 | 78.5 | 96.0 | 98.7 | 61.7 | 90.3 | 95.6 | 520.8 |
| SGRAF(Diao et al. 2021) | 77.8 | 94.1 | 97.4 | 58.5 | 83.0 | 88.8 | 499.6 | 79.6 | 96.2 | 98.5 | 63.2 | 90.7 | 96.1 | 524.3 |
| CMCAN(Zhang et al. 2022a) | 79.5 | 95.6 | 97.6 | 60.9 | 84.3 | 89.9 | 507.8 | 81.2 | 96.8 | 98.7 | **65.4** | 91.0 | 96.2 | 529.3 |
| NAAF†(Zhang et al. 2022b) | 81.9 | 96.1 | **98.3** | **61.0** | 85.3 | 90.6 | 513.2 | 80.5 | 96.5 | 98.8 | 64.1 | 90.7 | **96.5** | 527.2 |
| CHAN†*(Pan et al. 2023) | 79.7 | 94.5 | 97.3 | 60.2 | 85.3 | 90.7 | 507.8 | 79.7 | 96.7 | 98.7 | 63.8 | 90.4 | 95.8 | 525.0 |
| Set-Based†(Kim et al. 2023) | 80.9 | 94.7 | 97.6 | 59.4 | **85.6** | **91.1** | 509.3 | 80.6 | 96.3 | 98.8 | 64.7 | 91.4 | 96.2 | 528.0 |
| **NUIF-d***(ours) | 81.8 | 95.7 | 98.0 | 59.0 | 83.9 | 89.9 | 508.3 | 79.9 | 96.7 | 99.0 | 63.9 | 90.4 | 95.8 | 525.7 |
| **NUIF**(ours) | **84.3** | **96.3** | 98.0 | 60.7 | 85.0 | 90.7 | **515.1** | **81.7** | **97.0** | **99.0** | 65.1 | **91.4** | 96.3 | **530.6** |
| **BUTD + BERT** | | | | | | | | | | | | | | |
| GPO*(Chen et al. 2021) | 81.7 | 95.4 | 97.6 | 61.4 | 85.9 | 91.5 | 513.5 | 79.7 | 96.4 | 98.9 | 64.8 | 91.4 | 96.3 | 527.5 |
| VSRN++(Li et al. 2022a) | 79.2 | 94.6 | 97.5 | 60.6 | 85.6 | 91.4 | 508.9 | 77.9 | 96.0 | 98.5 | 64.1 | 91.0 | 96.1 | 523.6 |
| MV-VSE*(Li et al. 2022b) | 82.1 | 95.8 | 97.9 | 63.1 | 86.7 | 92.3 | 517.5 | 80.4 | 96.6 | 99.0 | 64.9 | 91.2 | 96.0 | 528.1 |
| CHAN*(Pan et al. 2023) | 80.6 | 96.1 | 97.8 | 63.9 | 87.5 | 92.6 | 518.5 | 81.4 | 96.9 | 98.9 | 66.5 | 92.1 | 96.7 | 532.6 |
| HREM(Fu et al. 2023) | 84.0 | 96.1 | 98.6 | 64.4 | 88.0 | 93.1 | 524.2 | 82.9 | 96.9 | 99.0 | 67.1 | 92.0 | 96.6 | 534.6 |
| **NUIF-d***(ours) | 83.9 | 96.5 | 98.2 | 67.9 | 89.2 | 93.6 | 529.4 | 83.3 | 97.3 | 98.9 | 69.2 | 92.7 | 96.9 | 538.2 |
| **NUIF**(ours) | **85.6** | **97.2** | **98.6** | **69.8** | **90.4** | **94.4** | **535.9** | **84.7** | **97.5** | **99.1** | **70.6** | **93.1** | **97.2** | **542.3** |

Table 1: Comparisons with state-of-the-arts on Flickr30K and MSCOCO 1K test sets. The † : the model has GloVe attached for text embedding, and $*$ : only single model is reported. The bests are in bold.

Flickr30K contains $31,000$ images and each image with $5$ texts. Following dataset splits in (Lee et al. 2018), $29,000$ images for training, $1,000$ images for validation, and $1,000$ images for testing. MSCOCO contains $133,287$ images and each image with $5$ texts. We use $123,287$ images for training, $5,000$ images for validation, and $5,000$ images for testing, and the results on MSCOCO are reported by both averaging over 5 folds of $1,000$ test images and testing on the entire $5,000$ test images. As common in the field of information retrieval, we measure the performance by recall R@K and rSum. The higher R@K indicates better performance.

## 5.2 Implementation Details

We utilize the BUTD features (Anderson et al. 2018) extracted from Faster R-CNN (Ren et al. 2015) with pre-trained ResNet-101 (He et al. 2016) as ROI inputs. $M = 36$ ROIs in each image. The dimension $D = 1024$ and $P = 256$. The temperature parameter $\lambda = 9.0$ and $\tau = 6.0$. The number of blocks $N = 16$. The margin $\alpha = 0.2$. In the semantic dependency gathering, we calculate polar coordinates $(\rho, \theta)$ of other regions relative to the target region and select regions with the first 2 small $\rho$ in each of the scopes that are quartered by $\theta = \pi/4, 3\pi/4, -3\pi/4, -\pi/4$, and extract noun phrases from texts by the chunking function of the NLP tool spaCy. In using Bi-GRU, the dropout operation is applied on both region and word features after projecting them into $1024$-dim and dropout rate is $0.4$, and we employ the Adam optimizer with $0.0002$ initial learning rate which is decayed by $10$ times after $40$ epochs on Flickr30K and after $30$ epochs on MSCOCO. In using BERT, the Adam optimizer sets $0.0005$ initial learning rate, and decays by $10$ times after $20$ epochs. Source code will be released[1].

[1]https://github.com/htzhang-code/NUIF

| Methods | IMG → TXT | | | TXT → IMG | | | rSum |
|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| **BUTD + Bi-GRU** | | | | | | | |
| SGRAF | 57.8 | - | 91.6 | 41.9 | - | 81.3 | - |
| CMCAN | 61.5 | - | 92.9 | **44.0** | - | 82.6 | - |
| NAAF† | 58.9 | 85.2 | 92.0 | 42.5 | 70.9 | 81.4 | 430.9 |
| CHAN†* | 60.2 | 85.9 | 92.4 | 41.7 | 71.5 | 81.7 | 433.4 |
| Set-Based† | 60.4 | 86.2 | 92.4 | 42.6 | **73.1** | **83.1** | 437.8 |
| **NUIF-d***(ours) | 59.3 | 85.5 | 92.0 | 41.9 | 71.3 | 81.8 | 431.8 |
| **NUIF**(ours) | **61.8** | **86.6** | **93.1** | 43.3 | 72.4 | 82.6 | **439.8** |
| **BUTD + BERT** | | | | | | | |
| VSRN++ | 54.7 | 82.9 | 90.9 | 42.0 | 72.2 | 82.7 | 425.4 |
| MV-VSE* | 59.1 | 86.3 | 92.5 | 42.5 | 72.8 | 83.1 | 436.3 |
| CHAN* | 59.8 | 87.2 | 93.3 | 44.9 | 74.5 | 84.2 | 443.9 |
| HREM | 64.0 | 88.5 | 93.7 | 45.4 | 75.1 | 84.3 | 450.9 |
| **NUIF-d***(ours) | 65.2 | 88.8 | 94.2 | 48.3 | 76.8 | 85.7 | 459.1 |
| **NUIF**(ours) | **67.8** | **89.8** | **94.8** | **49.9** | **77.9** | **86.7** | **466.9** |

Table 2: Comparisons with state-of-the-arts on MSCOCO 5K test set. The bests are in bold.

## 5.3 Comparisons with State-of-the-art Methods

We compare our proposed NUIF with recent state-of-the-art methods on the Flickr30K and MSCOCO benchmarks. The experimental results are cited directly from respective papers. When using the BUTD+Bi-GRU encoder, for a fair comparison with more previous methods, we report performances without the pre-trained GloVe representation attached to text embedding. Quantitative results on Flickr30K and COCO 1K test sets are shown in Tab. 1. NUIF outperforms state-of-the-art methods on most metrics with large margins clearly and achieves consistent superiority in different encoder settings. Comparisons on COCO 5k test set are shown in Tab. 2, and our method also performs best on al-

| Methods | IMG → TXT | | | TXT → IMG | | | rSum |
|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| **BUTD + Bi-GRU** | | | | | | | |
| w/o {dropout, $\mathrm{PN}_f$ } | 73.7 | 92.0 | 96.0 | 54.9 | 80.2 | 86.7 | 483.4 |
| w/o $\mathrm{PN}_f$ | 79.4 | 94.9 | 97.8 | 58.6 | 83.4 | 90.0 | 504.1 |
| NUIF-d (w/ $\mathrm{PN}_f$-$d$) | 81.8 | 95.7 | 98.0 | 59.0 | 83.9 | 89.9 | 508.3 |
| NUIF-r (w/ $\mathrm{PN}_f$-$r$) | 82.5 | 94.9 | 97.9 | 59.6 | 83.8 | 90.1 | 508.8 |
| NUIF-full | 84.3 | 96.3 | 98.0 | 60.7 | 85.0 | 90.7 | 515.1 |
| **BUTD + BERT** | | | | | | | |
| w/o $\mathrm{PN}_f$ | 82.0 | 95.9 | 98.5 | 66.8 | 88.6 | 93.4 | 525.2 |
| NUIF-d (w/ $\mathrm{PN}_f$-$d$) | 83.9 | 96.5 | 98.2 | 67.9 | 89.2 | 93.6 | 529.4 |
| NUIF-r (w/ $\mathrm{PN}_f$-$r$) | 83.7 | 96.5 | 98.4 | 67.4 | 89.1 | 93.3 | 528.4 |
| NUIF-full | 85.6 | 97.2 | 98.6 | 69.8 | 90.4 | 94.4 | 535.9 |

Table 3: Ablation studies of $\mathrm{PN}_f$'s modeling on Flickr30K.

| Methods | IMG → TXT | | | TXT → IMG | | | rSum |
|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| **COCO 1K, BUTD + Bi-GRU** | | | | | | | |
| w/o {dropout, $\mathrm{PN}_f$ } | 76.0 | 94.9 | 97.9 | 61.0 | 88.0 | 94.4 | 512.2 |
| w/o $\mathrm{PN}_f$ | 78.3 | 96.2 | 98.8 | 63.2 | 90.2 | 95.6 | 522.3 |
| NUIF-d (w/ $\mathrm{PN}_f$-$d$) | 79.9 | 96.7 | 99.0 | 63.9 | 90.4 | 95.8 | 525.7 |
| NUIF-r (w/ $\mathrm{PN}_f$-$r$) | 80.0 | 96.4 | 98.8 | 63.7 | 90.5 | 95.7 | 525.2 |
| NUIF-full | 81.7 | 97.0 | 99.0 | 65.1 | 91.4 | 96.3 | 530.6 |
| **COCO 5K, BUTD + Bi-GRU** | | | | | | | |
| w/o {dropout, $\mathrm{PN}_f$ } | 54.2 | 82.5 | 89.2 | 39.5 | 68.0 | 78.5 | 412.0 |
| w/o $\mathrm{PN}_f$ | 58.0 | 84.7 | 91.8 | 41.4 | 70.8 | 81.2 | 427.9 |
| NUIF-d (w/ $\mathrm{PN}_f$-$d$) | 59.3 | 85.5 | 92.0 | 41.9 | 71.3 | 81.8 | 431.8 |
| NUIF-r (w/ $\mathrm{PN}_f$-$r$) | 59.9 | 85.3 | 92.1 | 41.8 | 71.3 | 81.1 | 431.5 |
| NUIF-full | 61.8 | 86.6 | 93.1 | 43.3 | 72.4 | 82.6 | 439.8 |
| **COCO 1K, BUTD + BERT** | | | | | | | |
| w/o $\mathrm{PN}_f$ | 82.9 | 97.0 | 98.8 | 68.7 | 92.6 | 96.8 | 536.8 |
| NUIF-d (w/ $\mathrm{PN}_f$-$d$) | 83.3 | 97.3 | 98.9 | 69.2 | 92.7 | 96.9 | 538.2 |
| NUIF-r (w/ $\mathrm{PN}_f$-$r$) | 84.2 | 97.2 | 99.1 | 69.3 | 92.6 | 96.9 | 539.3 |
| NUIF-full | 84.7 | 97.5 | 99.1 | 70.6 | 93.1 | 97.2 | 542.3 |
| **COCO 5K, BUTD + BERT** | | | | | | | |
| w/o $\mathrm{PN}_f$ | 64.8 | 88.2 | 94.2 | 47.8 | 76.5 | 85.4 | 456.9 |
| NUIF-d (w/ $\mathrm{PN}_f$-$d$) | 65.2 | 88.8 | 94.2 | 48.3 | 76.8 | 85.7 | 459.1 |
| NUIF-r (w/ $\mathrm{PN}_f$-$r$) | 65.6 | 89.1 | 94.3 | 48.4 | 76.6 | 85.8 | 459.8 |
| NUIF-full | 67.8 | 89.8 | 94.8 | 49.9 | 77.9 | 86.7 | 466.9 |

Table 4: Ablation studies of $\mathrm{PN}_f$'s modeling on MSCOCO.

most all metrics. The remarkable improvements of our proposed NUIF demonstrate its effectiveness and robustness.

## 5.4 Ablation Study

To demonstrate that necessary semantic undertaker identifying does play an active role in accurately measuring image-text relevance, we conduct ablation studies on Flickr30K and MSCOCO for probability-of-necessity modeling, as enumerated in Tab. 3 and 4. The baseline w/o $\mathrm{PN}_f$ means to aggregate alignments without identifying necessity, *i.e.*, alignments averaging, and NUIF-full is ensemble model from NUIF-d and NUIF-r. In BUTD+Bi-GRU, dropout operation (see Sec. 5.2) is beneficial to improve baseline. It can be seen that our necessary undertakers identifying improves the matching accuracy significantly.

It is worth noting that MSCOCO's performance gains are less significant than Flickr30K's since MSCOCO's weaker causality (due to its much smaller proportion of not-so-short text) rather than scale (see Fig. 4). Short texts have weak causality since they are generally rough, while not-so-short ones have rich causality since their fine-grained details, *e.g.*, many regions in a sunset image are aligned with "beautiful sunset", removing some regions (*e.g.*, cloud) will not affect the degree of semantic sharing between sunset image and text, resulting in ambiguous (weak) causality. While for text "beautiful sunset with white clouds over a river", the necessity of certain regions (*e.g.*, cloud) is enhanced.

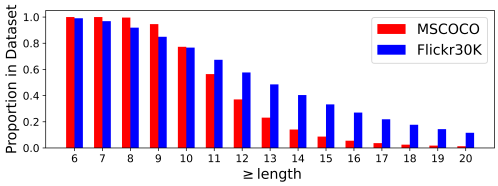



Figure 4: Illustration of the proportion of long text in MSCOCO is relatively smaller than that of Flickr30K.


Figure 5: Visualization of the probability-of-necessity. We highlight regions with large $\mathrm{PN}_i^{\mathrm{v}}$-$d$, and annotate the min-max scaled vaules of $\mathrm{PN}_j^{\mathrm{u}}$-$d$ for critical words.

## 5.5 Visualization

To further verify our method's ability to specify the necessary undertakers of the degree of semantic sharing between image and text, we visualize the learned $\mathrm{PN}_f$-$d$ of regions and words in Fig. 5. In row 2 column 2, due to the spatial proximity strategy in gathering region dependency, the girl closer to the fish region that semantically corresponds to the match-critical word "fish" is wrongly considered more necessary than the other. However, on the whole, our method effectively captures the regions and words necessary for judging semantic consistency or contradiction in matching.

## 6 Conclusion

In this paper, we revisit image-text matching in the causal view, and propose a novel theoretical prototype for estimating the probability-of-necessity of fragments for the degree of semantic sharing by means of counterfactual inference. Further, we implement a Necessary Undertaker Identification Framework (NUIF) for image-text matching to formalize the probability-of-necessity of fragments in two ways, which intuitively specifies the contribution of fragments to image-text relevance. Our method attributes the degree of image-text semantic sharing to constituent semantics. Extensive experiments demonstrate the superiority of our proposed NUIF. Future works include designing effective semantic dependency gathering, to reasonably infer fragments' necessity in specific scenarios.

## A Necessary Undertaker Identification

*Proof.* The semantic dependency $d$ of the fragment $f$ gathers up those fragments in image or text that have direct semantic causalities with $f$, containing $f$ itself. Removing $f$ will break these causalities and then distort the semantics emerged from dependency $d$. This is equivalent to the original semantics of $d$ being altered from the image or text. That is, as $o_f$ changes to $o'_f$, $o_d$ changes to $o'_d$ on semantics. Combining the definition of necessary cause (see Sec. 3.2) in causal inference (Pearl 2009; Glymour, Pearl, and Jewell 2016) , we express the probability that fragment $f$ is necessary to the degree of semantic sharing between image and text, probability-of-necessity, as:

$$\mathrm{PN}_f = P\left(m'_{o'_d} \mid m, o_d\right), \tag{13}$$

which means the probability that, given that $d$ did occur and $M$ = true in reality, the potential response of matching degree $M$ to $d$'s hypothetical erasure is $M$ = false. Since the matching degree $M$ is determined by image-text relevance $R$ monotonically and uniquely, and $D$ is exogenous relative to the relevance $R$, *i.e.*, $\left\{R_{o_d}, R_{o'_d}\right\} \perp\!\!\!\perp D$, then:

$$\left\{M_{o_d}, M_{o'_d}\right\} \perp\!\!\!\perp D, \tag{14}$$

which implies:

$$P(m_{o_d}) = P(m_{o_d}|o_d) = P(m|o_d), \tag{15}$$

that is:

$$o_d \wedge m = o_d \wedge m_{o_d}. \tag{16}$$

Then, for Eq. 13, we have:

$$\begin{aligned} P\left(m'_{o'_d} \mid m, o_d\right) &= \frac{P\left(m'_{o'_d}, m, o_d\right)}{P(m, o_d)} = \frac{P\left(m'_{o'_d}, m_{o_d}, o_d\right)}{P(m, o_d)} \\ &= \frac{P\left(m'_{o'_d}, m_{o_d}\right) P(o_d)}{P(m, o_d)} = \frac{P\left(m'_{o'_d}, m_{o_d}\right)}{P(m \mid o_d)}. \end{aligned} \tag{17}$$

Obviously, $m_{o'_d} \vee m'_{o'_d}$ = true, then:

$$\begin{aligned} m_{o_d} =& m_{o_d} \wedge \left(m_{o'_d} \vee m'_{o'_d}\right) \\ =& \left(m_{o_d} \wedge m_{o'_d}\right) \vee \left(m_{o_d} \wedge m'_{o'_d}\right), \end{aligned} \tag{18}$$

and $m_{o_d} \vee m'_{o_d}$ = true, then:

$$m_{o'_d} = \left(m_{o'_d} \wedge m_{o_d}\right) \vee \left(m_{o'_d} \wedge m'_{o_d}\right), \tag{19}$$

considering the matching monotonicity, $m_{o'_d} \wedge m'_{o_d}$ = false:

$$m_{o'_d} = m_{o'_d} \wedge m_{o_d}. \tag{20}$$

Substituting Eq. 20 into Eq. 18, it holds that:

$$m_{o_d} = m_{o'_d} \vee \left(m'_{o'_d} \wedge m_{o_d}\right). \tag{21}$$

Since the disjointness of $m_{o'_d}$ and $m'_{o'_d}$, and of $m_{o'_d}$ and $m_{o_d}$ (exogeneity of $d$), we obtain:

$$P(m_{o_d}) = P(m_{o'_d}) + P(m_{o_d}, m'_{o'_d}), \tag{22}$$

then taking the exogeneity of $d$, it yeilds:

$$P(m|o_d) = P(m|o'_d) + P(m_{o_d}, m'_{o'_d}). \tag{23}$$

Combining Eq. 17 and Eq. 23, we have:

$$P\left(m'_{o'_d} \mid m, o_d\right) = \left(P\left(m \mid o_d\right) - P\left(m \mid o'_d\right)\right) / P(m \mid o_d). \tag{24}$$

Thus, we obtain:

$$\mathrm{PN}_f = \left(P\left(m \mid o_d\right) - P\left(m \mid o'_d\right)\right) / P(m \mid o_d), \tag{25}$$

which concludes the proof. It is worth noting that, for a matched image-text pair, the matching degree $m$ ($M$ = true) means they are semantically related and $m'$ ($M$ = false) means the relationship of matched is no longer vaild. For unmatched image and text, the $m$ indicates the semantic relevance they achieve is being maintained and $m'$ indicates a weakening of the relevance. □

## Acknowledgements

This work is supported by the National Science Fund for Excellent Young Scholars under Grant 62222212.

## References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Chen, J.; Gao, Z.; Wu, X.; and Luo, J. 2023a. Meta-causal Learning for Single Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7683–7692.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the Best Pooling Strategy for Visual Semantic Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15789–15798.

Chen, L.; Zheng, Y.; Niu, Y.; Zhang, H.; and Xiao, J. 2023b. Counterfactual samples synthesizing and training for robust visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, T.; and Luo, J. 2020. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10583–10590.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity Reasoning and Filtration for Image-Text Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Fu, Z.; Mao, Z.; Song, Y.; and Zhang, Y. 2023. Learning Semantic Relationship Among Instances for Image-Text Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15159–15168.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ji, Z.; Chen, K.; and Wang, H. 2021. Step-Wise Hierarchical Alignment Network for Image-Text Matching. In *IJCAI*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kim, D.; Kim, N.; and Kwak, S. 2023. Improving Cross-Modal Retrieval With Set of Diverse Embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23422–23431.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 201–216.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2022a. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 641–656.

Li, Y.; Yang, X.; Shang, X.; and Chua, T.-S. 2021. Interventional video relation detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4091–4099.

Li, Z.; Guo, C.; Feng, Z.; Hwang, J.-N.; and Xue, X. 2022b. Multi-View Visual Semantic Embedding. In *IJCAI*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, C.; Mao, Z.; Zhang, T.; Xie, H.; Wang, B.; and Zhang, Y. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10921–10930.

Liu, R.; Liu, H.; Li, G.; Hou, H.; Yu, T.; and Yang, T. 2022. Contextual Debiasing for Visual Recognition With Causal Mechanisms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12755–12765.

Liu, Y.; Chen, J.; Chen, Z.; Deng, B.; Huang, J.; and Zhang, H. 2021. The blessings of unlabeled background in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6176–6185.

Lv, F.; Liang, J.; Li, S.; Zang, B.; Liu, C. H.; Wang, Z.; and Liu, D. 2022. Causality Inspired Representation Learning for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8046–8056.

Mao, C.; Xia, K.; Wang, J.; Wang, H.; Yang, J.; Bareinboim, E.; and Vondrick, C. 2022. Causal Transportability for Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7521–7531.

Pan, Z.; Wu, F.; and Zhang, B. 2023. Fine-Grained Image-Text Matching by Cross-Modal Hard Aligning Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19275–19284.

Pearl, J. 2009. *Causality*. Cambridge university press.

Qu, L.; Liu, M.; Wu, J.; Gao, Z.; and Nie, L. 2021. Dynamic modality interaction modeling for image-text retrieval. In *ACM SIGIR*, 1104–1113.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3716–3725.

Wang, H.; Zhang, Y.; Ji, Z.; Pang, Y.; and Ma, L. 2020a. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, 18–34. Springer.

Wang, S.; Wang, R.; Yao, Z.; Shan, S.; and Chen, X. 2020b. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1508–1517.

Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020c. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10760–10770.

Wehrmann, J.; Kolling, C.; and Barros, R. C. 2020. Adaptive cross-modal embeddings for image-text alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12313–12320.

Wei, H.; Wang, S.; Han, X.; Xue, Z.; Ma, B.; Wei, X.; and Wei, X. 2022. Synthesizing Counterfactual Samples for Effective Image-Text Matching. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4355–4364.

Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; and Wu, F. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10941–10950.

Yan, S.; Yu, L.; and Xie, Y. 2021. Discrete-continuous Action Space Policy Gradient-based Attention for Image-Text Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8096–8105.

Yang, X.; Feng, F.; Ji, W.; Wang, M.; and Chua, T.-S. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR*

*Conference on Research and Development in Information Retrieval*, 1–10.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *International Conference on Learning Representations*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. 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.

Zang, C.; Wang, H.; Pei, M.; and Liang, W. 2023. Discovering the Real Association: Multimodal Causal Reasoning in Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19027–19036.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020a. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, H.; Mao, Z.; Zhang, K.; and Zhang, Y. 2022a. Show Your Faith: Cross-Modal Confidence-Aware Network for Image-Text Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, K.; Mao, Z.; Wang, Q.; and Zhang, Y. 2022b. Negative-Aware Attention Framework for Image-Text Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15661–15670.

Zhang, K.; Zhang, L.; Hu, B.; Zhu, M.; and Mao, Z. 2023a. Unlocking the Power of Cross-Dimensional Semantic Dependency for Image-Text Matching. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4828–4837.

Zhang, Q.; Lei, Z.; Zhang, Z.; and Li, S. Z. 2020b. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3536–3545.

Zhang, S.; Song, X.; Li, W.; Bai, Y.; Yu, X.; and Jiang, S. 2023b. Layout-Based Causal Inference for Object Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10792–10802.