(Article begins on next page)

BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues

Sara Sarto[1], Marcella Cornia[1], Lorenzo Baraldi[1], and Rita Cucchiara[1,2]

[1] University of Modena and Reggio Emilia, Italy
[2] IIT-CNR, Italy
{name.surname}@unimore.it

Abstract. Effectively aligning with human judgment when evaluating machine-generated image captions represents a complex yet intriguing challenge. Existing evaluation metrics like CIDEr or CLIP-Score fall short in this regard as they do not take into account the corresponding image or lack the capability of encoding fine-grained details and penalizing hallucinations. To overcome these issues, in this paper, we propose BRIDGE, a new learnable and reference-free image captioning metric that employs a novel module to map visual features into dense vectors and integrates them into multi-modal pseudo-captions which are built during the evaluation process. This approach results in a multimodal metric that properly incorporates information from the input image without relying on reference captions, bridging the gap between human judgment and machine-generated image captions. Experiments spanning several datasets demonstrate that our proposal achieves state-of-the-art results compared to existing reference-free evaluation scores. Our source code and trained models are publicly available at: https://github.com/aimagelab/bridge-score.

Keywords: Captioning Evaluation · Vision-and-Language

1 Introduction

The objective of image captioning is to produce natural language descriptions conditioned on input images, that closely resemble human language and align to human intentions. As such, the captioning task involves the recognition and understanding of the visual content of the image, including fine-grained elements such as objects, attributes, and their relationships. Advances in training methodologies and architectures have contributed to the progress in the field, significantly improving the generation quality. Recent innovations include fully-attentive models [13, 14, 17], improved connections between visual and textual modalities [14, 41], and the incorporation of objects and tags at an architectural level [3, 31, 63]. Additionally, there has been a notable focus on increasing the robustness of cross-modal features [5, 32, 49], which consequently can increase description accuracy.

As constant improvements are made on the generation side, it becomes crucial to enhance the evaluation process as well. In this regard, image captioning
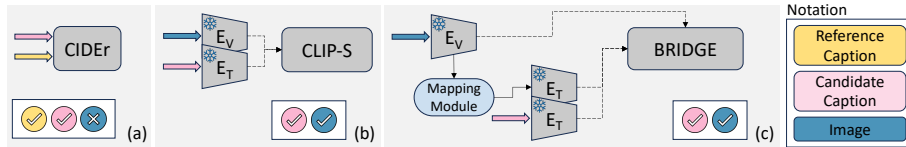
Fig. 1: Comparison between different captioning evaluation approaches: (a) CIDEr [56] scores candidate and reference captions without considering the input image; (b) CLIP-Score [18] compares text and images using global vectors in a shared embedding space; (c) our BRIDGE, internally builds multimodal pseudo-captions by translating fine-grained image features into pseudo-tokens thanks to a mapping module.

evaluation aims to assess the quality of a generated caption given an image and, potentially, human-written reference captions. However, it is important to note that obtaining these reference captions can often be challenging and expensive, adding complexity to the evaluation process. Despite the recent advancements in captioning capabilities, standard automatic evaluation metrics have mainly relied on translation metrics [4, 33, 42] or text-only ones [2, 56, 64] which often fall short in capturing aspects such as grammatical correctness, semantic relevance, and specificity. These limitations are worsened by the limited coverage of image content in available references, resulting in inaccurate penalties when generated captions accurately describe novel elements not mentioned in the references.

In response to these limitations, advanced metrics aligning visual and textual data have emerged [18, 24, 26, 27]. Notably, recent metrics leverage the CLIP embedding space [43], which shows a strong correlation with human judgment. Despite the effectiveness of contrastive-based embedding spaces, metrics based on dual-encoder architectures tend to focus on global alignment between an image and its caption, often ignoring fine-grained details or penalizing hallucinations.

Following this insight, in this paper, we introduce BRIDGE, a novel learnable and *reference-free* image captioning metric that enhances the alignment of more fine-grained visual features. Specifically, our model provides a pre-trained dual-encoder architecture with a mapping module designed to effectively exploit visual cues (Fig. 1). This is done by internally creating multimodal pseudo-captions, containing both textual and dense visual features. The process for building these pseudo-captions involves the creation of a template caption, which focuses on the syntactical structure of the scene, and a mapping module. The latter refines the template caption by enriching it with more fine-grained visual features about the subjects depicted in the image. Subsequently, the overall model is trained with a combination of contrastive losses which promote multimodal alignment.

Experiments are conducted on a variety of datasets for image captioning annotated with human rankings, including Flickr8k-Expert, Flickr8k-CF [19], Composite [1], and Pascal-50S [56]. Through comprehensive analysis, we demonstrate the efficacy of the proposed metric and show its ability to overcome the limitations of existing state-of-the-art reference-free alternatives. Additionally, we evaluate its sensitivity to object hallucination by conducting experiments on the FOIL dataset [50]. Overall, BRIDGE outperforms previous metrics and showcases superior performance compared to CLIP-Score [18] and PAC-Score [48].

Contributions. In summary, our contributions are as follows:
- We tackle the limitations of existing image captioning metrics by proposing the first learnable reference-free metric, termed as BRIDGE, that focuses on more fine-grained visual features. Our proposal integrates a dual-encoder architecture with a mapping module which is in charge of producing multimodal pseudo-captions that combine text and richer dense visual features.
- Multimodal pseudo-captions are built by creating template captions that describe the scene from a syntactical point of view. These templates are subsequently enriched with fine-grained features through a mapping network.
- Experiments, carried out on different datasets with human preferences, demonstrate a higher degree of correlation with respect to existing metrics. We also evaluate the sensitivity of the proposal to objects hallucinations.

2   Related Work

Classical Reference-based Captioning Metrics. Several widely used captioning evaluation metrics were originally developed in the context of NLP tasks and rely on n-gram matching techniques. Among these classical metrics, BLEU [42] is designed to focus on precision and incorporates a penalty for sentence brevity. METEOR [4], instead, combines precision and recall to evaluate the quality of captions, while others, such as ROUGE [33], were initially born for summarization tasks and later adapted to image captioning. More recently, two metrics tailored for visual captioning have emerged: CIDEr [56], which measures n-gram similarity and is based on TF-IDF, and SPICE [2], which quantifies graph-based similarity through scene graphs constructed from candidate and reference captions. Overall, focusing on textual-level comparisons, these metrics assume that human-written references accurately represent the image content.

Learnable Captioning Metrics. With the rise of large pre-trained models, image captioning evaluation now frequently exploits these models to compare textual-only [61, 64] or visual-textual [18, 21, 22, 26, 27, 58] contents. Notably, the BERT score and its improved version use pre-trained BERT embeddings to compare word tokens in generated and ground-truth sentences.

Some metrics, like BLEU and CIDEr, rely solely on text matching between reference captions and machine-generated captions, potentially introducing bias in evaluations due to non-accurate reference captions. To mitigate these issues, alternative solutions leverage the multimodal nature of vision-and-language models. As an example, TIGEr [22] considers the similarities between words and image regions and assesses how well machine-generated captions represent image content and their alignment with human-generated captions.

In contrast, other approaches [18, 24, 27] leverage web-scale vision-and-language models such as VilBERT [39] and CLIP [43] for more robust metrics. For example, in [24], CLIP visual-textual features are used to compute negative Gaussian cross-mutual information. Other works, instead, have employed diffusion models in text-only tasks [65] or exploited the zero-shot language modeling capabilities of large language models [8] to evaluate candidate captions.

Reference-free Captioning Metrics. While all aforementioned metrics rely on a set of ground-truth captions to compute the final score, a few attempts have been made to introduce reference-free evaluations, only taking into account the correlation of the candidate caption with the image. In this regard, Lee *et al.* [26] proposed to fine-tune the UNITER model [10] via contrastive learning to let the model to discriminate between positive and synthetically-generated negative captions. On a different line, Hessel *et al.* [18] introduced the CLIP-Score, which only relies on a modified cosine similarity between image and candidate caption representations coming from the CLIP model. The recently proposed PAC-Score [48], instead, builds upon the usage of CLIP but incorporates a fine-tuning phase with positive augmentation, further enhancing the accuracy of evaluation. In this paper, we follow this research path and propose a novel learnable, and reference-free evaluation metric that can effectively incorporate fine-grained visual features for evaluating the correlation of image-text pairs.

3   BRIDGE for Captioning Evaluation

In the following, we present our reference-free captioning evaluation approach, termed **BRIDGE**. Our approach leverages a dual-encoder architecture, which comprises both a language encoder and a vision encoder. Given a frozen pre-trained model, we train a mapping module responsible for filling the holes of a masked template caption with *pseudo language tokens* that are enriched with visual information. An overview of our model is depicted in Fig. 2.

3.1   Preliminaries

Our approach relies on CLIP (Contrastive Language–Image Pre-training) [43], a powerful vision and language model designed to align images and corresponding text captions within a shared embedding space. For a given input image I, the image encoder E_V extracts the visual information $v = E_V(I) \in \mathbb{R}^d$. On the textual side, an input caption T is tokenized and a textual representation is obtained by passing it through the textual encoder E_T, obtaining $t = E_T(T) \in \mathbb{R}^d$. Once the textual and visual features, t and v respectively, are projected in a common space, visual and textual inputs can be compared via cosine similarity.

The relationships learned by CLIP can be exploited to build an image captioning evaluator. In CLIP-Score [18] the authors directly compare candidate captions and images in the embedding space and show that this achieves a good correlation with human judgments. In detail, to assess the quality of a candidate generation, they feed both the image and the candidate caption through their respective feature extractors, and compute the cosine similarity of the resultant embeddings:

$$\text{CLIP-Score}(I, T) = w \cdot \max(\cos(v, t), 0), \tag{1}$$

where w is a rescaling factor employed to stretch the score distribution while ensuring the ranking results remain unchanged.
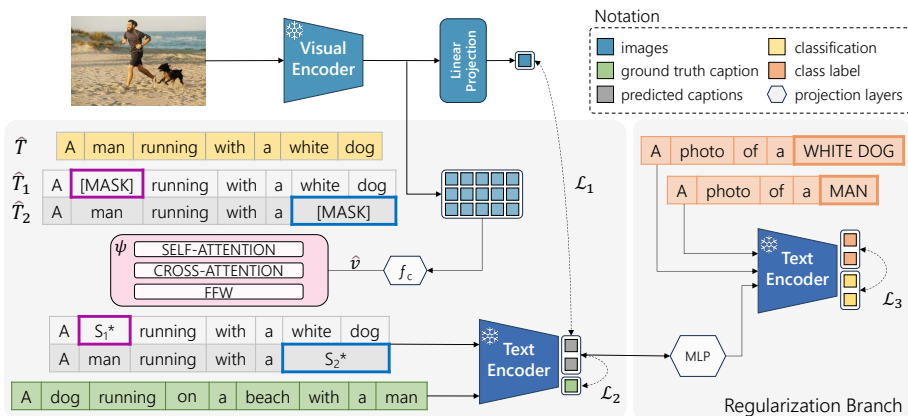
Fig. 2: Overview of the BRIDGE evaluation approach. Starting from a template caption, a mapping network augments it with dense visual features, obtaining a pseudo-caption which is then used for computing image-text similarities.

3.2 Injecting Fine-Grained Visual Features

Unlike CLIP-Score, our approach does not rely exclusively on global image descriptors for evaluating image-text alignments. Instead, we focus on employing stronger visual information. To do so, we draw inspiration from the Pic2Word approach [47] and represent the input image through a multimodal pseudo-caption, an embedding representation that contains stronger visual elements.

Building Template Captions. In order to create multimodal pseudo-captions, we first build *template captions* for a given input image. These are skeletal textual representations of the image, obtained by masking out all the relevant textual concepts from the descriptions generated by a captioner. Through these template captions, we aim to provide the model only with a templated textual structure which can then be filled with more fine-grained visual features. In particular, given an automatically generated caption describing the input image, such as `'A man running with a white dog'`, we remove the main subjects within the caption (*e.g.* `'man'` and `'white dog'`) and mask them with a `[MASK]` token. This will allow the model to fill in these gaps by incorporating more fine-grained features from the image encoder. Since a primary subject might be described by words other than just its corresponding nouns (*e.g.* adjectives), we utilize noun chunks. Fig. 3 reports template captions and corresponding noun chunks.

Given a sentence containing N noun chunks, we independently encode them through the mapping network. To this aim, we replicate the template caption as many times as the number of noun chunks and mask a different noun chunk in each of the replicas. We thus obtain N different versions of the template caption, each one masking a single noun chunk, for instance

```
['A [MASK] running with a white dog',
   'A man running with a [MASK]'].
```
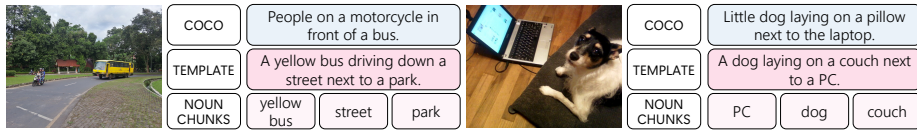
Fig. 3: COCO captions with template captions and associated noun chunks.

Mapping with Fine-Grained Visual Features. The above-described masked replicas are then fed to the mapping module ψ. Specifically, our approach exploits the visual information extracted from the visual encoder E_V to enrich the replicas with the informative content of the image I. To get a more fine-grained representation of the image, we directly take the grid of features from the last layer, $\hat{v}$. For instance, in the case of a ViT-B/32 backbone, this will have a shape of $50 \times d$, where d is the dimensionality of the last embedding of the network.

The mapping network is implemented as a stack of Transformer [55] encoder layers interleaved with cross-attention layers. Its role is to refine each template captions with visual information. Since each template caption is processed independently, the mapping module returns a set of sequences, each with the same length as the corresponding input template caption. From the output of the mapping module, we keep only the predictions for the masked tokens in each template caption and copy them back into the original templates.

Therefore, by providing a masked input template in the form $\hat{T}_i = [w_1, ..., w_{j-1}, \texttt{MASK}, w_{j+1}, ..., w_T]$, where $\{w_j\}_j$ represent original tokens from the input caption, we obtain $\hat{T_i}^* = \left[w_1, ..., w_{j-1}, \psi(\hat{T}_i)_j, w_{j+1}, ..., w_T\right]$, where $\psi(\hat{T}_i, \hat{v})_j$ represents the output of the mapping network at the position corresponding to the masked input token position. In the case of noun chunks consisting of more than one token, multiple consecutive tokens are replaced with the corresponding outputs from the mapping network. By injecting the outputs of the mapping token into the initial template caption, we effectively complete the original templates with visually enriched vectors. Notably, these newly generated pseudo-captions combine word sequences from the template captions with *dense vectors* obtained by the mapping module. Consequently, they cannot be decoded as standard captions. As a last step, the obtained pseudo-captions are fed into the pre-trained CLIP language encoder.

3.3   Training Protocol

To train our mapping network, the loss is defined as a weighted version of the symmetric InfoNCE loss [40], where positive and negative items are weighted according to the number of noun chunks in each caption.

Specifically, given a batch in the form $\mathcal{B} = \{(I_i, T_i)\}_{i=1}^N$, where I_i and T_i represent image-caption pairs, each image I_i is expanded in N_i multimodal pseudo-captions as outlined above, where N_i is the number of noun chunks in caption T_i. Further, let $\hat{t_{ij}}^*$ represent the embedding vector of the j-th pseudo-caption derived from the i-th image, v_i the embedding vector of the i-th image and t_i

the embedding vector of the i-th ground-truth caption. Finally, let M be the total number of noun chunks in the mini-batch, *i.e.* $M = \sum_{i=1}^{N} N_i$.

The first loss, denoted as $\mathcal{L}_1$, tries to align the pseudo-captions $\hat{t_{ij}}^*$ with the global visual features of the corresponding images v_i. In addition to this loss term, we define a second loss component $\mathcal{L}_2$ that promotes the alignment between pseudo-captions $\hat{t_{ij}}^*$ and the textual feature vector of the ground-truth caption t_i corresponding to the input image. This ensure that pseduo-captions are aligned also on a textual space, in addition to being aligned in the image space. Additional details can be found in the supplementary materials.

Regularization Branch. With the aforementioned loss terms, our objective is encouraging an association between each pseudo-caption and its corresponding image and caption. However, it is also important to differentiate each pseudo-caption from the others of the same image. To achieve this, we define a regularization loss which promotes a precise alignment between each pseudo-caption and the corresponding noun chunk.

First, we create prompts like "`a photo of a <NOUNCHUNK>`" and encode them with the text encoder E_T. In parallel, each pseudo-caption is fed into a dedicated multilayer perceptron (MLP) projection, which consist of two linear layers with a ReLU activation in between. Formally, the branch is defined as

$$\mathcal{C}(x) = \texttt{Linear}(\texttt{ReLU}(\texttt{Linear}(x))). \tag{2}$$

Since our goal is to emphasize each pseudo-caption's alignment with its corresponding noun chunk, we employ a regular contrastive loss $\mathcal{L}_3$ between the prompts mentioned earlier and the outputs of the projection branch.

Finally, the overall loss function we train BRIDGE is defined as a weighted summation of two aforementioned losses, plus the regularization loss, as

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_r. \tag{3}$$

3.4   Inference and Score Computation

At inference time, given an image-candidate caption pair (I, T), we extract all pseudo-captions from I using our mapping network. Subsequently, we compute the mean pseudo-caption embedding as $t^* = \frac{1}{N}\hat{t_i}^*$, where $\hat{t_i}^*$ indicates the i-th pseudo-caption extracted from I and N here indicates the overall number of pseudo-captions associated with I.

At that point, given the visual embedding v of the image and the embedding of the candidate caption t, the matching score between I and T is defined as

$$\text{BRIDGE}(I, T) = 0.5 \cdot [\text{CLIP-Score}(I, T) + w \cdot \max(\cos(t^*, t), 0)], \tag{4}$$

where cos indicates the cosine similarity and w is a constant scaling factor.

4   Experimental Evaluation

4.1   Implementation Details

Architecture and Training Details. Building upon prior research [18,24,51], we use either CLIP [43] ViT-B/32 or ViT-L/14 as backbone for the visual and textual encoder. The mapping module is composed of two Transformer layers and is trained on the COCO dataset [34], which contains more than 120k images annotated with five captions. In particular, we employ the splits introduced by Karpathy *et al.* [23], where 5,000 images are used for both validation and testing and the rest for training. To map the grid visual features to an embedding space of dimension 512, we employ a simple linear projection. For the regularization branch, we utilize a two-layer multi-layer perceptron.

During training, we use AdamW [38] as optimizer with a learning rate equal to 0.0001 and a batch size of 256. The λ_1, λ_2, and λ_3 values are selected with a grid search, choosing the combination that provides the best validation loss. Specifically, we set both λ_1 and λ_3 to 0.01, while λ_2 is set to 1.0. The training stage lasts around one day on a single A100 GPU.

Template Caption Generation. The template captions used as input for the mapping module are generated using the BLIP model [29]. In particular, we use the ViT-L/14 version pre-trained on 129M image-text pairs and finetuned on the COCO dataset. After this generation phase, the primary subjects of the template sentences are extracted by using the NLTK library [6]. During training, two noun chunks are randomly chosen from the set identified during the extraction step. In the evaluation phase, otherwise, all identified noun chunks are included.

4.2   Datasets

To evaluate the correlation of the proposed metric with human ratings, we conduct experiments on the Flickr8k-Expert, Flickr8k-CF, Composite, and Pascal50-S datasets [1,19,56]. In addition, for detecting hallucinations in textual sentences, we extend our analysis to the FOIL dataset [50]. Except for Pascal-50S and FOIL in where accuracy scores are used, evaluation on all other datasets relies on Kendall τ_b, Kendall τ_c, and Spearman ρ correlation scores.

Flickr8k-Expert and Flickr8k-CF [19]. These datasets consist of image-caption pairs with corresponding human ratings. Specifically, Flickr8k-Expert comprises 17k expert annotations for visual-textual pairs, with a total of 5,664 different images. Each pair receives a score ranging from 1 (lack of correlation) to 4 (accurate depiction), where 1 indicates a lack of correlation between the caption and the image, and 4 indicates an accurate depiction of the image without errors. On the other hand, Flickr8k-CF is composed of 145k binary quality judgments, collected from CrowdFlower, for 48k image-caption pairs containing 1,000 unique images. Each pair is annotated with at least three binary scores, where "yes" denotes that the caption correlates with the image. To measure the alignment with human judgment, we compute the mean proportion of "yes" annotations as the score for each pair.

Table 1: Ablation study results. US indicates the number of pseudo tokens.

	Flickr8k-Expert			Flickr8k-CF			Pascal-50S
	Kend. τ_b	Kend. τ_c	Spear. ρ	Kend. τ_b	Kend. τ_c	Spear. ρ	Accuracy
Architectural Components							
w/o mapping module (*i.e.* MLP)	53.1	53.5	65.3	35.5	18.3	43.5	81.9
w/o template captions	53.7	54.1	66.0	35.5	18.4	43.6	82.5
w/o regularization branch	54.1	54.5	66.5	35.7	18.5	43.6	82.7
Score Formulation							
w/o textual similarity	51.1	51.2	63.0	34.4	17.7	30.5	80.9
w/o visual similarity	53.8	54.2	66.0	35.4	18.3	43.7	81.9
Pseudo-token Size							
w/ US = 1	54.1	54.5	66.4	35.1	17.1	30.3	82.6
w/ US = 2	54.1	54.5	66.4	**36.1**	**18.7**	44.4	**82.8**
w/ US = 4	54.3	54.7	66.6	35.4	18.3	43.8	82.4
w/ US = 8	54.0	54.4	66.3	35.9	18.6	**44.5**	81.7
BRIDGE (US = 3)	**54.4**	**54.8**	**67.7**	**36.1**	**18.7**	**44.5**	82.6

Composite [1]. It contains 12k human ratings for image-caption pairs including a combination of images taken from COCO [34] (2,007 images), Flickr8k [19] (997 images), and Flickr30k [62] (991 images). In this dataset, human evaluators assess each image-caption pair, assigning a score within the range of 1 to 5 to estimate the correspondence of the caption with the associated image.

Pascal50-S [56]. It presents pairwise preference judgments between two captions. Overall, the dataset consists of 4,000 sentence pairs, each of them associated with an image from the UIUC Pascal sentence dataset [46]. Each pair is associated with 48 human judgments, where each evaluation indicates which sentence better describes the given image. The sentence pairs are categorized into four groups: (i) both human-written and correct captions (HC), (ii) both human-written captions where one is correct and the other is wrong (HI), (iii) both correct captions but one written by humans and the other machine-generated (HM), (iv) both machine-generated and correct captions (MM).

FOIL [50]. The dataset comprises image-caption pairs from the COCO dataset [34]. In this scenario, captions undergo perturbation by generating modified versions that closely resemble the originals but introduce a single error, referred to as "foil word". For a fair comparison, we select the subset of the validation set that does not overlap with the portion of COCO used during training, resulting in 8,000 images, each paired with a foil-correct textual counterpart.

4.3 Ablation Studies and Analysis

To evaluate the effectiveness of our metric, we start by analyzing variations of our main architectural components. Then, we assess the impact of caption templates in our score formulation. All these experiments are performed using CLIP ViT-B/32 as backbone and reported in Table 1.

Contribution of Architectural Components. We first investigate the performance of the most straightforward implementation of a mapping module, structured as a two-layer MLP following [47]. We also validate the importance

Table 2: Impact of different template captions.

	Expert	CF	Pascal-50S
	Kend. τ_b	Kend. τ_b	Acc.
Transformer Templates			
w/o mapping module	46.1	31.6	80.4
BRIDGE	**54.0** (+7.9)	**35.9** (+4.3)	**82.7** (+2.3)
BLIP Templates			
w/o mapping module	48.6	33.8	81.0
BRIDGE	**54.4** (+5.8)	**36.1** (+2.3)	**82.6** (+1.6)
BLIP-2 Templates			
w/o mapping module	49.4	34.2	82.4
BRIDGE	**54.4** (+5.0)	**36.2** (+2.0)	**82.9** (+0.5)

COCO — A young girl is preparing to blow out her candle / A painting of a table with fruit on top of it

TRANSF — A young girl is eating a piece of cake / A painting of a painting of fruits and a painting

BLIP — A woman blowing out a candle on a cake / A painting of a pitcher and oranges on a table

BLIP-2 — A girl blowing out a candle on a cupcake / A painting of a still life with oranges and a candle

of the template captions through a model variant in which a set of learnable tokens S^* serves as input for the mapping module, without relying on template captions. In both variants, given the absence of template captions, we construct a template such as 'a photo of S^*' and extract its features using the CLIP text encoder. In the Table, it can be seen that, regardless of any architectural changes, it is important to provide a simple sentence structure to the mapping module to achieve competitive performance.

In addition to these baselines, we devise a variant to analyze the contribution of the regularization branch. In this setting, we employ template captions as input for the mapping module, resulting in a substantial improvement of $+1.0$ Kendall τ_b and $+0.8$ accuracy points compared to the MLP variant, respectively on the Flickr8k-Expert and on the Pascal-50S dataset. When introducing the regularization branch (*i.e.* the complete BRIDGE architecture), further enhancements can be observed especially on the Flickr8k-CF with an improvement of $+0.4$ points in terms of the Kendall τ_b correlation score.

We also emphasize the importance of each component in our score formulation. Specifically, we present correlation results when employing only the visual similarity within our architecture, which is the original CLIP-Score formulation. As observed, performance drops drastically when relying only on visual information. A less significant drop is observed when employing only textual similarity.

As an additional analysis, we report the effect of changing the number of the pseudo tokens for each noun chunk, denoting it as unit size (US). Specifically, we compute the scores employing $US = 1, 2, 3, 4, 8$. From the results, it can be seen that $US = 3$ generally leads to the best performance across nearly all evaluation metrics. This configuration is used in all experiments reported in the paper.

Analysis on Caption Templates. We analyze the effect of changing the initial template captions for our model. Specifically, we employ template captions generated by a conventional Transformer-based captioner that uses CLIP features as input and is trained only on the COCO dataset, as well as those generated by BLIP [29], and BLIP-2 [28]. Notably, we select template captions of different quality based on both standard metric evaluations and correlations with human judgment. To assess the quality of the raw generated templates, we observe that

the CIDEr score of these models on the COCO test set is equal to 114.2, 131.4, and 145.8 respectively for the standard Transformer model, BLIP, and BLIP-2. Note that all models were trained with cross-entropy loss only. Results on the Flickr8k-Expert, Flickr8k-CF, and Pascal-50S datasets are reported in Table 2. To qualitatively validate the generated templates, we include sample captions generated by the three models compared to a ground-truth caption from the COCO test set. Specifically, captions generated by BLIP-2 are generally more detailed and effectively describe the visual content of the input image compared to those generated by BLIP and, notably, the standard Transformer model.

For each caption template source, we compute the correlation scores when the mapping module is disabled and the captions are directly fed to the text encoder. We compare it with our standard BRIDGE score, considering the different caption templates. Across all datasets, it is evident that directly using the caption templates as input to the text encoder leads to poor performance. This highlights the intended flexibility of template captions as skeletal representations, allowing the model to enhance them with fine-grained visual features.

In fact, starting from these simple template captions and following our approach, we achieve improvements of +5.8 and +2.3 Kendall τ_b points and +1.6 accuracy points, respectively, on the Flickr8k-Expert, Flickr8k-CF, and Pascal50-S datasets when using BLIP caption templates. The overall best results are with captions from the BLIP-2 model, confirming that better templates can indeed lead to improved results. However, even when using lower-quality captions, the final correlation results are very close to those obtained with higher-quality captions. This highlights the robustness of our metric to caption templates of varying quality and that it is not necessary to rely on captions generated by large-scale captioners to achieve strong correlation scores.

4.4 Comparison with State-of-the-Art Captioning Metrics

Evaluating Sample-Level Human Correlation. We evaluate the sample-level human correlation on the Flickr8k [19] and Composite [1] datasets. Following previous works [18,64], we compute Kendall correlation scores in both τ_b and τ_c versions and also include the Spearman ρ score. Results are reported in Table 3, where we compare our proposed BRIDGE metric against other reference-free evaluation scores like UMIC [26], CLIP-S [18], and PAC-S [48]. Moreover, we also compare with standard captioning evaluation metrics (*i.e.* BLEU [42], METEOR [4], CIDEr [56], and SPICE [2]) and more recent reference-based solutions that exploit text-only or cross-modal learned embeddings, such as BERT-S [64], BERT-S++ [61], TIGEr [22], VilBERTScore [27], and MID [24]. For completeness, we also include the reference-based versions of CLIP-S and PAC-S, termed RefCLIP-S and RefPAC-S, which however are not directly comparable with our solution as both rely on a set of five reference captions.

As it can be seen, BRIDGE outperforms other reference-free metrics in terms of correlation with human judgment, achieving the highest scores on almost all datasets. Specifically, compared to CLIP-S, BRIDGE shows improvements in terms of Kendall τ_c of +3.6 and +2.8 points when using ViT-B/32 and ViT-L/14

Table 3: Correlation scores on Flickr8k-Expert, Flickr8k-CF, and Composite [1,19].

	Flickr8k-Expert			Flickr8k-CF			Composite		
	Kend. τ_b	Kend. τ_c	Spear. ρ	Kend. τ_b	Kend. τ_c	Spear. ρ	Kend. τ_b	Kend. τ_c	Spear. ρ
Reference-based metrics									
BLEU-4 [42]	30.6	30.8	38.7	16.9	8.7	21.0	28.3	30.6	38.1
METEOR [4]	41.5	41.8	51.9	22.2	11.5	27.6	36.0	38.9	48.1
CIDEr [56]	43.6	43.9	54.3	24.6	12.7	30.5	34.9	37.7	47.0
SPICE [2]	51.7	44.9	55.1	24.4	12.0	31.3	38.8	40.3	49.1
BERT-S [64]	-	39.2	50.3	22.8	-	-	39.9	30.1	48.6
BERT-S++ [61]	48.1	46.7	56.9	-	-	-	42.3	44.9	52.1
TIGEr [22]	51.4	49.3	48.6	-	-	-	47.5	45.4	55.3
ViLBERTScore [27]	54.2	50.1	61.3	-	-	-	51.4	52.4	58.7
MID [24]	-	54.9	-	37.3	-	-	-	-	-
RefCLIP-S [18]	52.6	53.0	64.6	36.4	18.8	44.7	51.2	55.4	66.2
RefPAC-S [48]	55.5	55.9	67.6	37.6	19.5	46.2	53.0	57.3	68.0
Reference-free metrics									
UMIC [26]	-	46.8	-	-	-	-	-	-	-
CLIP-S (ViT-B/32) [18]	51.1	51.2	63.0	34.4	17.7	30.5	49.8	53.8	64.3
PAC-S (ViT-B/32) [48]	53.9	54.3	66.1	36.0	18.6	44.4	**51.5**	**55.7**	**66.3**
BRIDGE (ViT-B/32)	**54.4**	**54.8**	**66.7**	**36.1**	**18.7**	**44.5**	50.9	55.0	65.4
CLIP-S (ViT-L/14) [18]	52.6	53.0	64.7	35.2	18.2	43.3	51.3	55.4	65.9
PAC-S (ViT-L/14) [48]	55.1	55.5	67.3	**36.8**	19.0	**45.3**	52.3	56.5	67.1
BRIDGE (ViT-L/14)	**55.4**	**55.8**	**67.7**	36.3	**19.0**	44.7	**52.9**	**57.2**	**67.8**

backbone on the Flickr8k-Expert dataset. These improvements extend consistently across all datasets and correlation scores. Compared to PAC-S, BRIDGE still achieves superior results on both Flickr8k-Expert and Flickr8k-CF, while performing on par on the Composite dataset (*i.e.* PAC-S achieves the best results when using ViT-B/32, while BRIDGE outperforms PAC-S using ViT-L/14).

Discriminating Correct Captions. We also evaluate the effectiveness of our metric on the PASCAL-50S dataset [56]. In this context, instead of calculating correlation scores, we compute accuracy by determining, for each pair, the caption favored by the majority of human ratings as correct (with ties being resolved randomly). We then measure how frequently the evaluation metric assigns a higher score to the chosen caption. Following previous works [18], we randomly select five reference captions from the set of 48 provided by the dataset and average the results over five distinct draws. Accuracy values are reported in Table 4. The results show that when using

Table 4: Accuracy results on Pascal-50S [56] averaged over five random draws of reference captions. The † marker indicates scores from previous works.

|	HC	HI	HM	MM	Mean
Reference-based metrics					
BLEU-4 [42]	60.3	93.1	85.7	57.0	74.0
METEOR [4]	66.0	97.7	94.0	66.6	81.1
CIDEr [56]	66.5	97.9	90.7	65.2	80.1
BERT-S++† [61]	65.4	98.1	96.4	60.3	80.1
TIGEr† [22]	56.0	99.8	92.8	74.2	80.7
MID† [24]	67.0	99.7	97.4	76.8	85.2
RefCLIP-S [18]	64.9	99.5	95.5	73.3	83.3
RefPAC-S [49]	67.7	99.6	96.0	75.6	84.7
Reference-free metrics					
CLIP-S (ViT-B/32) [18]	55.9	99.3	96.5	72.0	80.9
PAC-S (ViT-B/32) [48]	**60.6**	99.3	96.9	72.9	82.4
BRIDGE (ViT-B/32)	59.4	**99.4**	**97.5**	**74.0**	**82.6**
CLIP-S (ViT-L/14) [18]	57.0	**99.6**	**96.7**	73.5	81.7
PAC-S (ViT-L/14) [48]	59.5	99.4	95.8	**74.7**	82.2
BRIDGE (ViT-L/14)	**61.2**	**99.6**	96.6	74.1	**82.9**
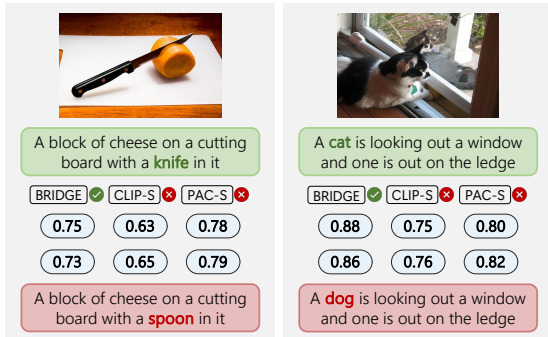
Fig. 4: Sample images from the FOIL dataset [50] and corresponding scores generated by our proposed metric compared with CLIP-S and PAC-S.

Table 5: Accuracy results on the FOIL [50] dataset.

	Acc.
Reference-based metrics	
BLEU-4 [42]	66.2
METEOR [4]	70.1
CIDEr [56]	85.7
MID [24]	90.5
RefCLIP-S [18]	91.0
RefPAC-S [18]	93.7
Reference-free metrics	
CLIP-S (ViT-B/32) [18]	87.2
PAC-S (ViT-B/32) [48]	89.9
BRIDGE (ViT-B/32)	**91.5**
CLIP-S (ViT-L/14) [18]	90.9
PAC-S (ViT-L/14) [48]	91.9
BRIDGE (ViT-L/14)	**93.0**

both ViT-B/32 and ViT-L/14, BRIDGE consistently outperform CLIP-S across all categories, showcasing an average accuracy increase of +1.7 and +1.2 points, respectively. When comparing with PAC-S, our solution can better discriminate the correct captions on both backbones with an average accuracy increase of +0.2 and +0.7 using ViT-B/32 and ViT-L/14 respectively.

Object Hallucination Analysis. We then extend our analysis to the FOIL dataset [50] for correctly identifying captions that may contain object hallucinations. Table 5 shows the accuracy results. As it can be seen, BRIDGE outperforms CLIP-S and PAC-S, exhibiting an increase respectively of +4.3 and +1.6 points when using ViT-B/32 as backbone. Similar improvements can also be observed when employing more robust visual features. These results demonstrate the capabilities of our metric also to identify hallucinated objects correctly. In Fig 4, we present sample results comparing our metric with CLIP-S and PAC-S.

4.5   System-level Correlation

Finally, we delve into the efficacy of our proposed metric when evaluating popular existing captioning models. To this aim, we generate predictions of several state-of-the-art captioning models on the COCO test set, including Show and Tell and Show, Attend and Tell which are among the first image captioning models based on deep learning, Up-Down [3], SGAE [60], AoANet [20], $\mathcal{M}^2$ Transformer [14], X-Transformer [41] which all include region-based image features with either LSTM-based or Transformer-based language models, and the recently proposed COS-Net model [32] that incorporates CLIP features. In addition to reporting evaluations on traditional captioning models, we also include recent LLM-based captioning models, including ZeroCap [52] and SmallCap [45], which are based on GPT-2 [44], and MiniGPT-v2 [9], BLIP-2 [28], IDEFICS [25], LLaVA-1.5 [35,36], and InstructBLIP [15] which instead are based on larger-scale LLMs like Flan-T5 [12], Vicuna [11], or LLaMA [53,54].

Table 6: Evaluation scores of traditional and LLM-based captioners on COCO test set (♦: not trained/fine-tuned on COCO).

		BLEU-4	METEOR	CIDEr	CLIP-S	PAC-S	**BRIDGE**
Traditional	Show and Tell [57]	31.4	25.0	97.2	0.572	0.772	0.788
	Show, Attend and Tell [59]	33.4	26.2	104.6	0.582	0.785	0.804
	Up-Down [3]	36.7	27.9	122.7	0.723	0.803	0.821
	SGAE [60]	39.0	28.4	129.1	0.734	0.812	0.833
	AoANet [20]	38.9	29.2	129.8	0.737	0.815	0.836
	$\mathcal{M}^2$ Transformer [14]	39.1	29.2	131.2	0.734	0.813	0.841
	X-Transformer [41]	39.7	29.5	132.8	0.610	0.812	0.845
	COS-Net [32]	42.0	30.6	141.1	0.758	0.832	0.859
LLM-based	ZeroCap♦ [52]	2.3	10.1	15.1	0.810	0.816	0.862
	SmallCap [45]	37.0	27.9	119.7	0.748	0.826	0.847
	MiniGPT-v2 [9]	18.8	24.6	80.4	0.752	0.818	0.845
	BLIP-2 [28]	**43.7**	**32.0**	**145.8**	0.767	**0.837**	0.868
	IDEFICS-9B♦ [25]	4.3	19.1	50.0	0.740	0.786	0.838
	LLaVA-1.5-7B♦ [35]	8.1	28.0	69.6	0.784	0.809	0.867
	InstructBLIP-Flan-T5-XL♦ [15]	6.1	28.1	38.1	**0.817**	**0.837**	**0.902**
	Humans	-	*24.1*	*87.6*	*0.774*	*0.823*	*0.856*

The results are presented in Table 6, where we evaluate our BRIDGE scores against both standard metrics, such as BLEU-4, METEOR, and CIDEr, and more recent ones like CLIP-S and PAC-S. Captioning models are compared to a human baseline, in which, for each sample, one human-annotated sentence (selected randomly from the five provided by the COCO dataset) serves as a candidate caption. As shown in the table, BRIDGE can effectively evaluate human-annotated sentences which obtain a score similar to recent state-of-the-art captioning models such as COS-Net. This capability lacks in standard metrics such as METEOR and CIDEr which rank human captions lower than those generated by less-performing captioning models like Show, Attend, and Tell or Up-Down.

When considering the evaluation of captions generated by existing models, our metric shows a strong correlation with standard evaluation metrics when evaluating traditional image captioners or large-scale models that are fine-tuned on the COCO dataset like BLIP-2. When instead considering more recent approaches that are based on large language models and are not fine-tuned on COCO, BRIDGE still recognizes the goodness of generated captions, raking InstructBLIP as the best-performing approach. This demonstrates the capabilities of our metric to correctly evaluate longer and more detailed captions which are typically generated by LLM-based multimodal models [7, 16, 30] and that might be very different from captions contained in the COCO dataset.

5   Conclusion

In this paper, we have presented a novel learnable, and reference-free image captioning metric that combines text and dense visual features. Our proposal, BRIDGE, employs templated captions that are enriched with fine-grained visual cues thanks to a mapping network. Through experimental evaluation, we demonstrate that BRIDGE outperforms existing reference-free metrics in terms of correlation with human judgment and sensitivity to hallucinated objects.

Acknowledgments

References

1. Aditya, S., Yang, Y., Baral, C., Fermuller, C., Aloimonos, Y.: From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge. arXiv preprint arXiv:1511.03292 (2015)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: Semantic Propositional Image Caption Evaluation. In: ECCV (2016)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
4. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshops (2005)
5. Barraco, M., Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: With a Little Help from your own Past: Prototypical Memory Networks for Image Captioning. In: ICCV (2023)
6. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc. (2009)
7. Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Baraldi, L., Cornia, M., Cucchiara, R.: The Revolution of Multimodal Large Language Models: A Survey. In: ACL Findings (2024)
8. Chan, D., Petryk, S., Gonzalez, J.E., Darrell, T., Canny, J.: CLAIR: Evaluating Image Captions with Large Language Models. In: EMNLP (2023)
9. Chen, J., Li, D.Z.X.S.X., Zhang, Z.L.P., Xiong, R.K.V.C.Y., Elhoseiny, M.: MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-task Learning. arXiv preprint arXiv:2310.09478 (2023)
10. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: UNiversal Image-TExt Representation Learning. In: ECCV (2020)
11. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality (2023)
12. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling Instruction-Finetuned Language Models. JMLR **25**(70), 1–53 (2024)
13. Cornia, M., Baraldi, L., Cucchiara, R.: SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability. In: ICRA (2020)
14. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-Memory Transformer for Image Captioning. In: CVPR (2020)

15. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv preprint arXiv:2305.06500 (2023)
16. Dong, H., Li, J., Wu, B., Wang, J., Zhang, Y., Guo, H.: Benchmarking and Improving Detail Image Caption. arXiv preprint arXiv:2405.19092 (2024)
17. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning: Transforming objects into words. In: NeurIPS (2019)
18. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP (2021)
19. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. JAIR **47**, 853–899 (2013)
20. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on Attention for Image Captioning. In: ICCV (2019)
21. Jiang, M., Hu, J., Huang, Q., Zhang, L., Diesner, J., Gao, J.: REO-Relevance, Extraness, Omission: A Fine-grained Evaluation for Image Captioning. In: EMNLP (2019)
22. Jiang, M., Huang, Q., Zhang, L., Wang, X., Zhang, P., Gan, Z., Diesner, J., Gao, J.: TIGEr: Text-to-Image Grounding for Image Caption Evaluation. In: EMNLP (2019)
23. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
24. Kim, J.H., Kim, Y., Lee, J., Yoo, K.M., Lee, S.W.: Mutual Information Divergence: A Unified Metric for Multimodal Generative Models. In: NeurIPS (2022)
25. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In: NeurIPS (2023)
26. Lee, H., Yoon, S., Dernoncourt, F., Bui, T., Jung, K.: UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. In: ACL (2021)
27. Lee, H., Yoon, S., Dernoncourt, F., Kim, D.S., Bui, T., Jung, K.: ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT. In: EMNLP Workshops (2020)
28. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: ICML (2023)
29. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: ICML (2022)
30. Li, X., Tu, H., Hui, M., Wang, Z., Zhao, B., Xiao, J., Ren, S., Mei, J., Liu, Q., Zheng, H., et al.: What If We Recaption Billions of Web Images with LLaMA-3? arXiv preprint arXiv:2406.08478 (2024)
31. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV (2020)
32. Li, Y., Pan, Y., Yao, T., Mei, T.: Comprehending and ordering semantics for image captioning. In: CVPR (2022)
33. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: ACL Workshops (2004)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV (2014)
35. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: CVPR (2024)

36. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: NeurIPS (2023)
37. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv preprint arXiv:2303.05499 (2023)
38. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: ICLR (2019)
39. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In: NeurIPS (2019)
40. Oord, A.v.d., Li, Y., Vinyals, O.: Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748 (2018)
41. Pan, Y., Yao, T., Li, Y., Mei, T.: X-Linear Attention Networks for Image Captioning. In: CVPR (2020)
42. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL (2002)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021)
44. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
45. Ramos, R., Martins, B., Elliott, D., Kementchedjhieva, Y.: SmallCap: lightweight image captioning prompted with retrieval augmentation. In: CVPR (2023)
46. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting Image Annotations Using Amazon's Mechanical Turk. In: NAACL Workshops (2010)
47. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2Word: Mapping Pictures to Words for Zero-Shot Composed Image Retrieval. In: CVPR (2023)
48. Sarto, S., Barraco, M., Cornia, M., Baraldi, L., Cucchiara, R.: Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In: CVPR (2023)
49. Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: Retrieval-Augmented Transformer for Image Captioning. In: CBMI (2022)
50. Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., Bernardi, R.: FOIL it! Find One mismatch between Image and Language caption. In: ACL (2017)
51. Shi, Y., Yang, X., Xu, H., Yuan, C., Li, B., Hu, W., Zha, Z.J.: EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching. In: CVPR (2022)
52. Tewel, Y., Shalev, Y., Schwartz, I., Wolf, L.: Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In: CVPR (2022)
53. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971 (2023)
54. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
56. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: Consensus-based Image Description Evaluation. In: CVPR (2015)

57. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR (2015)
58. Wang, S., Yao, Z., Wang, R., Wu, Z., Chen, X.: FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation. In: CVPR (2021)
59. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
60. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-Encoding Scene Graphs for Image Captioning. In: CVPR (2019)
61. Yi, Y., Deng, H., Hu, J.: Improving Image Captioning Evaluation by Considerg Inter Refinerences Variance. In: ACL (2020)
62. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL **2**, 67–78 (2014)
63. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: VinVL: Revisiting visual representations in vision-language models. In: CVPR (2021)
64. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. In: ICLR (2020)
65. Zhu, W., Wang, X.E., Yan, A., Eckstein, M., Wang, W.Y.: ImaginE: An Imagination-Based Automatic Evaluation Metric for Natural Language Generation. In: EACL (2023)

Supplementary Material

In the following, we present further experiments and analyses about the proposed BRIDGE metric. Specifically, we provide a detailed description of the weighted contrastive loss function used to train our approach. Additionally, we report several supplementary qualitative results to support our findings.

A   Weighted Contrastive Loss

In Section 3.2 of the main paper, we state that we employ a weighted variant of the symmetric InfoNCE loss [40]. Specifically, our method involves building mini-batches of multimodal pseudo-captions derived from a set of image-caption pairs. To recall the notation of that section, the mini-batched are in the form $\mathcal{B} = \{(I_i, T_i)\}_{i=1}^N$, where I_i and T_i represent image-caption pairs. Each image I_i is expanded in N_i multimodal pseudo-captions, with N_i representing the number of noun chunks in caption T_i. As in the main paper, we denote $\hat{t_{ij}}^*$ as the embedding vector of the j-th pseudo-caption derived from the i-th image, v_i as the embedding vector of the i-th image, and t_i as the embedding vector of the i-th ground-truth caption. Finally, let M be the total number of noun chunks in the mini-batch, *i.e.* $M = \sum_{i=1}^N N_i$.

The first weighted contrastive loss aligns the embeddings of the multimodal pseudo-captions with the global visual features of the corresponding images. This step ensures that each pseudo-caption is appropriately contextualized within the overall visual context. This loss is defined as a weighted version of the symmetric InfoNCE loss because positive and negative items are weighted according to the number of noun chunks in each caption. The rationale behind this choice is that captions having more noun chunks tend to have more visual variance. We therefore assign them a higher weight to promote the transfer of proper visual features. Formally, the loss is defined as follows

$$\mathcal{L}_1 = -\frac{1}{M} \sum_{i=1}^N \sum_{j=1}^{N_i} \log \frac{\exp(\cos(v_i, \hat{t_{ij}}^*)/\tau)}{\sum_{k \neq i} N_k \cdot \exp(\cos(v_k, \hat{t_{ij}}^*)/\tau)} +$$
$$-\frac{1}{M} \sum_{i=1}^N \sum_{j=1}^{N_i} N_i \cdot \log \frac{\exp(\cos(v_i, \hat{t_{ij}}^*)/\tau)}{\sum_{k \neq i} \sum_{h=1}^{N_k} \exp(\cos(v_i, \hat{t_{hk}}^*)/\tau)}. \tag{5}$$

Noticeably, differently from the standard InfoNCE loss, we also remove the positive item from the denominator of each loss component.

In addition to the above-defined loss, we define a second loss component that promotes the alignment between pseudo-captions and the textual feature vector of the ground-truth caption corresponding to the input image. This makes sure that pseudo-captions are aligned also on a textual space, in addition to being

Table 7: Human correlation and accuracy scores changing the underlying backbone.

	Expert	CF	Pascal-50S	FOIL
	Kendall τ_b	Kendall τ_b	Accuracy	Accuracy
CLIP-based backbone				
CLIP-S [18]	51.1	34.4	80.9	87.2
BRIDGE	**54.4**	**36.1**	**82.6**	**91.5**
PAC-based backbone				
PAC-S [48]	53.9	36.0	82.4	89.9
BRIDGE	**54.8**	**36.4**	**82.7**	**91.4**

aligned in the image space. Symmetrically to Eq. 5, this loss is defined as

$$\mathcal{L}_2 = -\frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{N_i} \log \frac{\exp(\cos(t_i, \hat{t_{ij}}^*)/\tau)}{\sum_{k \neq i} N_k \cdot \exp(\cos(t_k, \hat{t_{ij}}^*)/\tau)} +$$

$$-\frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{N_i} N_i \cdot \log \frac{\exp(\cos(t_i, \hat{t_{ij}}^*)/\tau)}{\sum_{k \neq i} \sum_{h=1}^{N_k} \exp(\cos(t_i, \hat{t_{hk}}^*)/\tau)}, \tag{6}$$

where t_i is the global textual feature vector of the ground-truth caption for V_i.

B Additional Experimental Results

Effect of Changing the Underling Backbone. The proposed BRIDGE score is based on the standard CLIP model without fine-tuning its original weights. However, recent solutions like PAC-S [48] improve the performance of CLIP-S by fine-tuning the final projections of visual and textual encoders with curated data. In Table 7, we assess whether applying the proposed BRIDGE approach to the fine-tuned backbone employed in PAC-S can further improve its final results. Interestingly, BRIDGE can not only enhance the results of a standard CLIP-based model but can also achieve improved correlation with human judgment when using the fine-tuned CLIP model presented in [48], termed as PAC in the table. This further demonstrates the effectiveness of our evaluation score and its generalization capabilities when employing different backbones.

Additional Ablation Studies. In the upper part of Table 8, we present the results across different datasets when employing the standard contrastive loss instead of considering the number of noun chunks in each caption. The results indicate that employing the standard loss does not enhance the final performance, thereby confirming the advantages of prioritizing captions that contain a greater number of noun chunks.

In the bottom part of Table 8, we report the results ablating other architectural choices. In particular, instead of taking the output of the mapping module in the correspondence of the `[MASK]` tokens, we feed the entire output to the textual encoder. Employing this model variant, referred to as "w/ entire mapping module output", when computing the BRIDGE metric leads to significantly lower correlation scores.

Table 8: Additional ablation study results on architectural design.

	Expert	CF	Pascal-50S
	Kendall τ_b	Kendall τ_b	Accuracy
Loss Design			
w/ standard contrastive losses	54.2	35.6	82.5
Architectural Design			
w/ entire mapping module output	52.0	34.5	82.5
w/ global features	53.9	35.1	82.3
BRIDGE	**54.4**	**36.1**	**82.6**
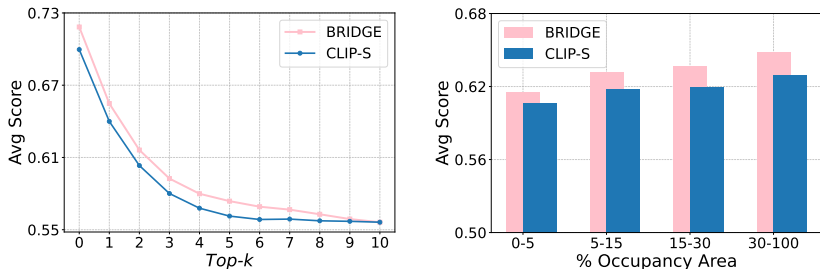
Fig. 5: Metric scores for top-k detections ranked by probability (left) and as a function of detection area (right).

We also investigate the impact of using global image features instead of grid-level features. In BRIDGE, we opt for employing grid-level features because the mapping module integrates visual features into the pseudo-caption, necessitating high-quality and detailed visual encoder outputs to effectively enhance the pseudo-captions. Indeed, upon examining the results, we observe a drop in the performance when using global visual features instead of more fine-grained grid features. This not only reinforces our choice but also emphasizes that our model can capture more robust fine-grained details by leveraging grid-level features.

This observation is also supported by the analysis reported in Fig. 5. In particular, we extract object detections using the Grounding DINO model [37] and compute scores between the image and the prompted class name ("`a photo of a <class>`") of each detected object. To ensure a fair comparison, we adjust BRIDGE distribution to match that of CLIP-S. As it can be seen, BRIDGE demonstrates higher confidence for larger objects. However, compared to CLIP, it assigns higher scores to all detections, indicating greater confidence even in smaller objects. Moreover, in the FOIL dataset, where a single detail is changed in the caption, BRIDGE demonstrates its stronger fine-grained capability by identifying hallucinated objects better than CLIP-S and PAC-S (cf. Table 5 of the main paper).

Qualitative Results. To qualitatively validate generated templates, we report in Fig. 6 additional sample captions generated by the three considered models in comparison to a ground-truth caption from the COCO test set. In particular, captions generated by BLIP-2 are generally more detailed and better describe the visual content of the input image compared to those generated by BLIP and, especially, the standard Transformer model.

COCO	A colored motorcycle rests outside of a sheep farm	A wooden dining table is surrounded by window	A herd of sheep standing on top of snow covered field	Baked pizza with red tomatoes and green olives
TRANSF	A motorcycle is parked in a desert field	A yellow flower in a vase on a table	A group of sheep grazing in a field	A pizza with tomatoes onions peppers and peppers
BLIP	A motorcycle parked in a dirt field next to a fence	A dining room table with a vase of sunflowers on it	A herd of sheep standing on top of a lush green field	A pizza sitting on top of a white plate
BLIP-2	A motorcycle parked in a dirt field with sheep in the back	A wooden table with a vase of sunflowers on it	A herd of sheep standing in a field next to a house	A pizza with tomatoes, olives and cheese on a white plate

Fig. 6: Sample captions generated by a standard Transformer model, BLIP, and BLIP-2, in comparison with ground-truth textual sentences from the COCO test set.

In Fig. 7, we present qualitative results on the FOIL dataset. We report results comparing BRIDGE to CLIP-S and PAC-S, showing that our proposed metric achieves better results in terms of detection of hallucinated objects. Additional comparisons of our metric with CLIP-S and PAC-s on the Pascal50-S dataset are presented in Fig. 8. Observing the results, although PAC-S, in some instances, aligns with human judgment, CLIP-S consistently assigns a lower score to the caption preferred by humans. On the other hand, BRIDGE metric demonstrates its effectiveness across the majority of cases.
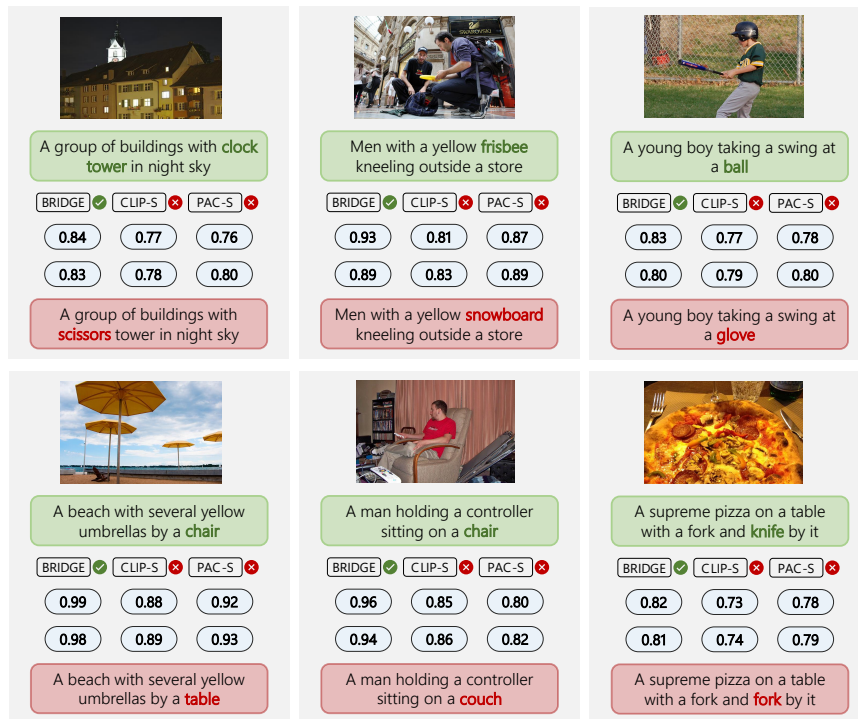
Fig. 7: Sample images from the FOIL hallucination detection dataset and corresponding evaluation scores generated by the BRIDGE metric in comparison with CLIP-S and PAC-S. Hallucinated objects are highlighted in red.
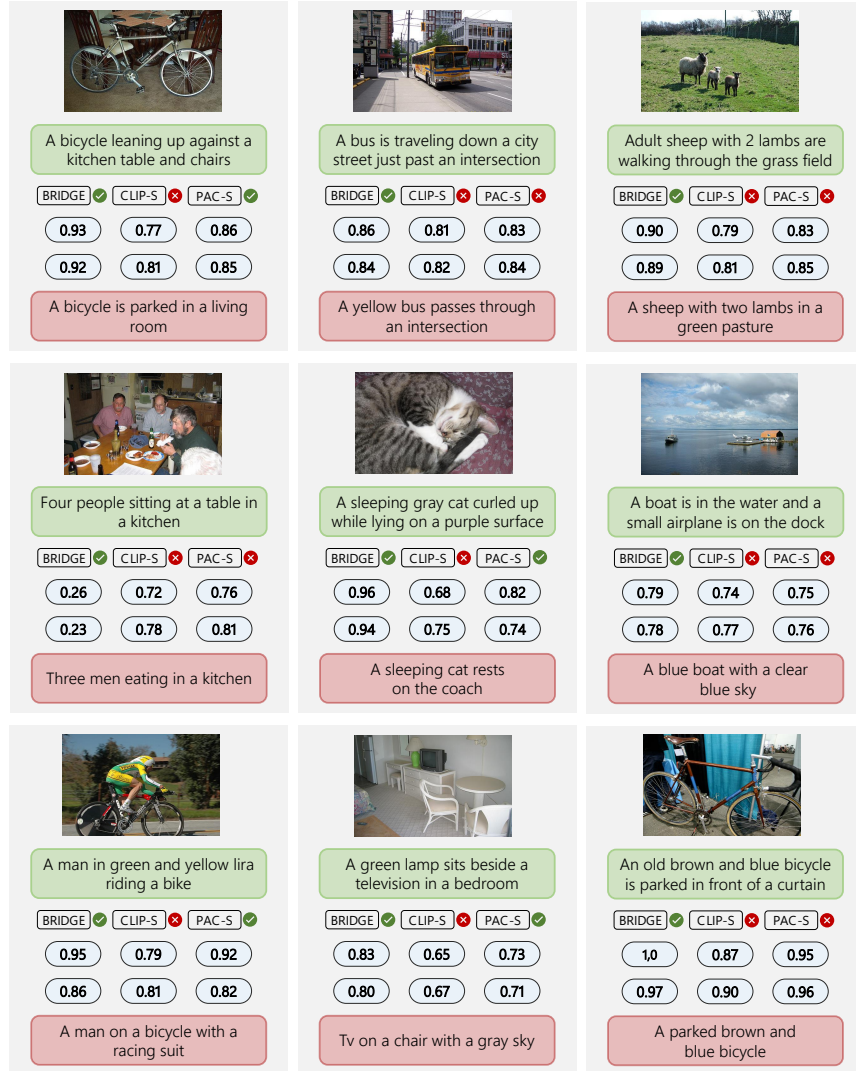
Fig. 8: Comparisons of recent metrics for captioning with respect to BRIDGE on the Pascal-50S dataset. The candidate caption in green is the one preferred by humans.