CROSS²STRA: Unpaired Cross-lingual Image Captioning with <u>Cross</u>-lingual <u>Cross</u>-modal <u>Str</u>ucture-pivoted <u>A</u>lignment

Shengqiong Wu, Hao Fei,* Wei Ji, Tat-Seng Chua
Sea-NExT Joint Lab, School of Computing, National University of Singapore
swu@u.nus.edu {haofei37, jiwei, dcscts}@nus.edu.sg,

Abstract

Unpaired cross-lingual image captioning has long suffered from irrelevancy and disfluency issues, due to the inconsistencies of the semantic scene and syntax attributes during transfer. In this work, we propose to address the above problems by incorporating the scene graph (SG) structures and the syntactic constituency (SC) trees. Our captioner contains the *semantic structure-guided image-to-pivot captioning* and the *syntactic structure-guided pivot-to-target translation*, two of which are joined via pivot language. We then take the SG and SC structures as pivoting, performing cross-modal semantic structure alignment and cross-lingual syntactic structure alignment learning. We further introduce cross-lingual&cross-modal back-translation training to fully align the captioning and translation stages. Experiments on English↔Chinese transfers show that our model shows great superiority in improving captioning relevancy and fluency.

1 Introduction

Generating texts to describe images (a.k.a., image captioning) has many real-world applications, such as virtual assistants and image indexing (Fang et al., 2015). Current image captioning models have achieved impressive performance (Jia et al., 2015; Gu et al., 2018a; Ji et al., 2021), yet are mainly limited to the English language due to the large-scale paired image-caption datasets. Subject to the scarcity of paired captioning data, the development of captioning in other languages is thus greatly hindered. While manually crafting sufficient paired data is prohibitively expensive, cross-lingual image captioning (Miyazaki and Shimizu, 2016) offers a promising solution, which aims to transfer a captioner trained at resource-rich language (e.g., English) to the resource-scarce language(s) without paired captioning data at target language(s).

A direct approach is to make use of the current translation techniques, i.e., the pivot language translation method. Here pivot language is the resource-rich language, e.g., English. For example, the pivot-side captioner first generates pivot captions for images, which are then translated into the target-side captions. Or one can create the pseudo image-caption pairs for directly training a target-side captioner, by translating the pivot training captions into the target ones (Lan et al., 2017). However, the above translation-based method suffers from two major issues (cf. 1(a)), including *irrelevancy* and *disfluency* (Song et al., 2019). On the one hand, due to the lack of paired vision contexts, a translated description can easily deviate from the original visual semantics, leading to ambiguous or inaccurate captioning. On the other hand, restricted to the translation system itself, translated texts often suffer from disfluent language, especially for the lengthy and complex descriptions.

Some previous efforts are carried out to rectify the above two key errors for better cross-lingual captioning. Lan et al. (2017) solve the translation disfluency issue by estimating the fluency of translation texts, then rejecting those disfluent ones. Yet their method dramatically sacrifices the paired training data, and meanwhile suffers from low-efficiency owing to the incremental screening process. Song et al. (2019) propose to enhance the relevance and fluency of translations by designing some rewards via the reinforcement learning technique. However, the *REINFORCE* algorithm (Williams, 1992) is hard to train, and easily leads to unstable results. We note that there are two critical abilities a cross-lingual captioning system should possess to solve the corresponding problems. For content relevancy, the kernel lies in sufficiently modeling the vision-language semantic alignment; while for language fluency, it is key to effectively capture the gaps of linguistic attributes and characteristics between the pivot and target languages.
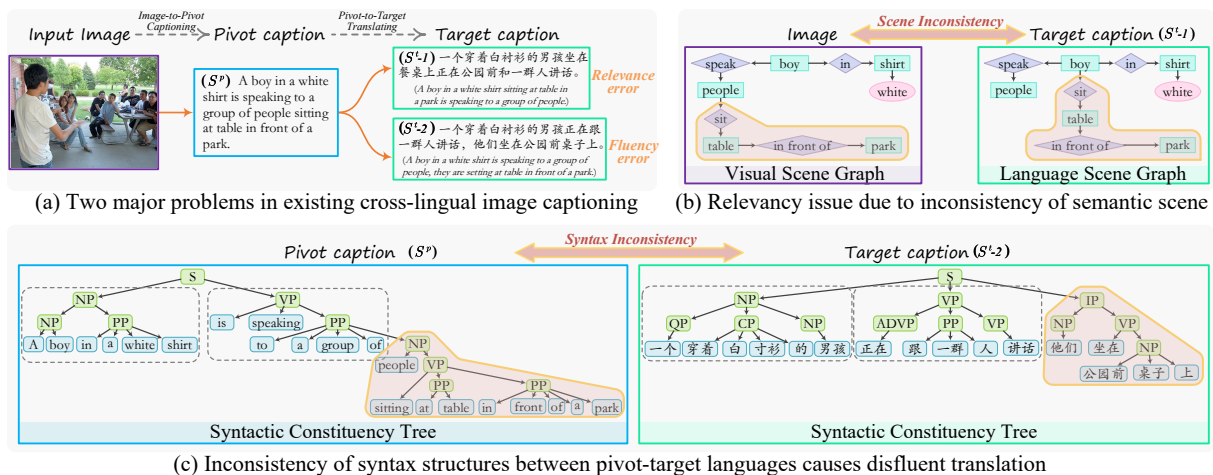
---

*Corresponding author: Hao Fei

Figure 1: We illustrate two issues in translation-based cross-lingual captioning (a), caused by the inconsistencies of semantic scene structures (b) and syntax structures (c). We simplify the constituency trees for better understanding, and the dotted grey box areas indicate the counterpart constituents across different languages. (Best viewed in color)

Besides the translation-based methods, the pivoting-based cross-lingual captioning methods have shown effectiveness, where the whole task learning is broken down into two steps, image-to-pivot captioning and pivot-to-target translation (Gu et al., 2018b; Gao et al., 2022). The image-to-pivot captioning learns to describe images in the pivot language based on pivot-side paired captioning data, and the pivot-to-target translation is performed based on parallel sentences. Two cross-model and cross-lingual subtasks are trained on two separate datasets, and aligned by the pivot language. Although achieving improved task performances, existing pivoting-based methods (Gu et al., 2018b; Gao et al., 2022) still fail to fully address the two major problems of cross-lingual captioning, due to the insufficient alignment of either vision-language semantics or pivot-target syntax.

To this end, we present a novel syntactic and semantic structure-guided model for cross-lingual image captioning. We build the framework based on the pivoting-based scheme, as shown in Fig. 2. For image-to-pivot captioning, we consider leveraging the scene graphs (SG) for better image-text alignment. Intuitively, an SG (Johnson et al., 2015; Yang et al., 2019) depicts the intrinsic semantic structures of texts or images, which can ideally bridge the gaps between modalities. For the pivot-to-target translating, we make use of the syntactic constituency (SC) tree structures for better pivot-target language alignment. Syntax features have been shown as effective supervisions for enhancing the translation quality, e.g., fluency and grammar-correctness (Schwartz et al., 2011; Xu et al., 2020;

Li et al., 2021).

Based on the above framework, we further perform cross-lingual cross-modal structure-pivoted alignment learning. First of all, we introduce an SG-pivoted cross-modal semantic structure alignment. Based on contrastive learning (Logeswaran and Lee, 2018; Yan et al., 2021) we realize the unsupervised vision-language semantic structure alignment, relieving the scene inconsistency and thus enhancing the relevancy. Similarly, an unsupervised SC-based cross-lingual syntax structure aligning is used to learn the shared grammar transformation and thus mitigate the language disfluency during translation. Finally, we perform the cross-lingual cross-modal back-translation training, fully aligning the two phrases of image-to-pivot captioning and pivot-to-target translation.

On English→Chinese and Chinese→English transfers of unpaired cross-lingual image captioning, our method achieves significant improvement over the existing best-performing methods. Further in-depth analyses demonstrate that the integration of both scene graph and syntactic structure features is complementarily helpful in improving the captioning relevancy and disfluency of the transfer.

Our main contributions are two-fold:

• First, we for the first time enhance the cross-lingual image captioning by leveraging both the semantic scene graph and the syntactic constituent structure information, such that we effectively address the problems of content irrelevancy and language disfluency.

• Second, we propose several cross-lingual cross-modal structure-pivoted alignment learning strate-

gies, via which we achieve effective cross-modal vision-language semantic alignment and cross-lingual pivot-target syntactic alignment.

2 Related Work

Image captioning has been an emerging task in the past few years and received great research attention (You et al., 2016; Vinyals et al., 2017; Cornia et al., 2020). Later, the task of cross-lingual image captioning (Miyazaki and Shimizu, 2016; Song et al., 2019) has been presented, to transfer the knowledge from resource-rich language to resource-poor language[1], so as to spare the burden of manual data annotation for the minority languages. However, the task has been hindered and received limited attention due to two key issues: irrelevancy and disfluency of captions. There are two categories of cross-lingual captioning approaches: the translation-based (Lan et al., 2017; Gu et al., 2018b) and the pivoting-based (Gu et al., 2018b; Gao et al., 2022) methods. The former employs an off-the-shelf translator to translate the source (pivot) captions into the target language for target-side training or as the target-side captions. The latter reduces the noise introduction of the pipeline by jointly performing the image-to-pivot captioning step and pivot-to-target translation step, thus being the current SoTA paradigm. This work inherits the success of this line, and adopts the pivoting-based scheme as a backbone, but we further strengthen it by leveraging the semantic and syntactic structure information to better solve the two issues.

Scene graphs depict the intrinsic semantic scene structures of images or texts (Krishna et al., 2017; Wang et al., 2018). In SGs, the key object and attribute nodes are connected to describe the semantic contexts, which have been shown useful as auxiliary features for wide ranges of downstream applications, e.g., image retrieval (Johnson et al., 2015), image generation (Johnson et al., 2018) and image captioning (Yang et al., 2019). Here we incorporate both the visual and language scene graphs to enhance the cross-modal alignment learning.

Note that Gao et al. (2022) also leverage the SG features for cross-lingual captioning, while ours differs from theirs in three aspects. First, they consider a fully unsupervised cross-lingual setup with no image-caption pairs at pivot language, while under such an unpaired assumption the visual and
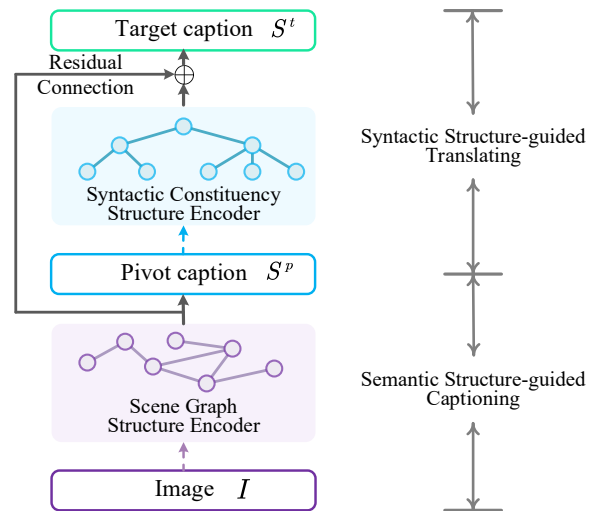

Figure 2: A high-level schematic of the cross-lingual captioning model architecture.

language scene graphs are hard to align, and thus limits the utility of SGs. Second, in this work we sufficiently align the two cross-modal SGs via unsupervised learning, such that the noises in SGs will be effectively screened. Third, Gao et al. (2022) align the pivot and target languages with also the SG structure. We note that it could be ineffective to perform cross-lingual alignment based on textual SGs because the scene structures in different languages are essentially the same.

In fact, two languages can be different the most in linguistic structures. Almost all the erroneous sentences come with certain grammar or syntax errors (Jamshid Lou et al., 2019, 2020). Also syntax features have been extensively found to be effective in improving the language quality (e.g., fluency and grammatically-correctness) in cross-lingual scenario (Nivre, 2015; Li et al., 2021; Zhang and Li, 2022). For example, in machine translation, different languages show great correspondences in phrasal constituent structures (Zhang and Zong, 2013; Fang and Feng, 2022). Also, syntactic structure features have been integrated into a broad number of downstream applications (Wu et al., 2021; Fei et al., 2021, 2022). Thus we consider making use of the syntax structures as cross-lingual supervision to enhance the captioning quality.

3 Syntactic Semantic Structure-guided Cross-lingual Captioning Framework

The original task is to learn a mapping $\mathcal{F}_{I \to S^t}$ from input images I to target-language captions S^t. Following Gu et al. (2018b); Song et al. (2019), we decompose $\mathcal{F}_{I \to S^t}$ into two mappings: 1) the

---

[1] Without using target-side image-caption pairs, the task is also known as unpaired cross-lingual image captioning.

image-to-pivot captioning $\mathcal{F}_{I \to S^p}$ training with the paired data $\{(I, S^p)\}$, and 2) the pivot-to-target translation $\mathcal{F}_{S^p \to S^t}$ training with the parallel data $\{(S^p, S^t)\}$. Note that $\{(I, S^p)\}$ and $\{(S^p, S^t)\}$ are two distinct datasets with possibly no intersection.

In our setting, we also leverage the SG and SC structure features in two mappings. As shown in Fig. 2, the semantic structure-guided captioning phase ($\mathcal{F}_{<I,\mathrm{SG}> \to S^p}$) takes as input the image I and the visual SG encoded by a structure encoder, yielding the pivot caption S^p. Then, the syntactic structure-guided translating phase ($\mathcal{F}_{<S^p,\mathrm{SC}> \to S^t}$) takes as input the S^p and the pivot SC, finally producing the target caption S^t. Note that the input embeddings of the second step are shared with the output embeddings from the first step so as to avoid the isolation of the two parts. Also we impose a residual connection from the SG feature representations to the SC feature representations to supervise the final target captioning with scene features.

3.1 Semantic Structure-guided Captioning

Given an image, we obtain its SG from an off-the-shelf SG parser, which is detailed in the experiment setup. We denote an SG as $SG=(V, E)$, where V is the set of nodes $v_i \in V$ (including object, attribute and relation types),[2] E is the set of edges $e_{i,j}$ between any pair of nodes v_i. We encode a SG with a graph convolution network (GCN; Marcheggiani and Titov, 2017):

$$\{\boldsymbol{h}_i\} = \mathrm{GCN}^G(SG), \qquad (1)$$

where $\boldsymbol{h}_i$ is the representation of a node v_i. We then use a Transformer (Vaswani et al., 2017) decoder to predict the pivot caption $\hat{S}^p$ based on $\{\boldsymbol{h}_i\}$:

$$\hat{S}^p = \mathrm{Trm}^G(\{\boldsymbol{h}_i\}). \qquad (2)$$

3.2 Syntactic Structure-guided Translation

In this step we first transform the predicted pivot caption S^p into the SC structure, $SC=(V, E)$, where V are the phrasal&word nodes connected by the compositional edge E. Different from the dependency-like SG structure, SC is a tree-like hierarchical structure, as depicted in Fig. 1. Similarly, we encode SC trees with another GCN:

$$\{\boldsymbol{r}_j\} = \mathrm{GCN}^C(SC), \qquad (3)$$

where $\boldsymbol{r}_j$ is an SC node representation. Another Transformer decoder is used to predict the target caption $\hat{S}^t$. To ensure the relevancy of target-side generation, we create a shortcut between the prior SG feature representations $\boldsymbol{h}$ and the SC features

---
[2] Appendix §A.1 details the SG and SC structures.
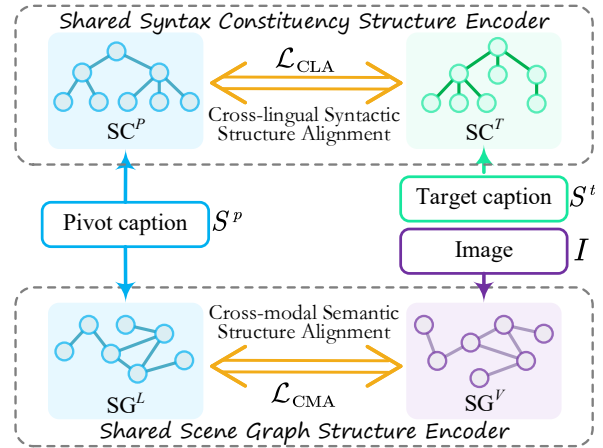

Figure 3: Illustration of the cross-modal semantic and cross-lingual syntactic structure alignment learning.

$\boldsymbol{r}$, via the cross-attention mechanism:

$$\hat{S}^t = \mathrm{Trm}^C(\{\boldsymbol{r}_j\}; \{\boldsymbol{h}_i\}). \qquad (4)$$

3.3 Two Separate Supervised Learning

The captioning and the translation steps are performed separately based on $\{(I, S^p)\}$ and $\{(S^p, S^t)\}$ in a supervised manner:

$$\mathcal{L}_{\mathrm{Cap}} = -\sum \log P(S^p | I, \mathrm{SG}), \qquad (5)$$
$$\mathcal{L}_{\mathrm{Tran}} = -\sum \log P(S^t | S^p, \mathrm{SC}). \qquad (6)$$

4 Structure-Pivoting Cross-lingual Cross-modal Alignment Learning

In the above supervised training, though leveraging the semantic and syntactic structure information, the cross-modal image-text pair and the cross-lingual pivot-target pair can be still under-aligned in their own feature spaces, due to the intrinsic structural gaps, e.g., noisy substructures. To combat that, we further propose two structure-pivoting unsupervised learning strategies (cf. Fig. 3): cross-modal semantic structure alignment and cross-lingual syntactic structure alignment. Besides, the two parts of our backbone captioner are initially trained separately. This motivates us to further align the two procedures in a whole-scale way, with cross-lingual&cross-modal back-translation training (cf. Fig. 4).

4.1 Cross-modal Semantic Structure Aligning

The basic idea is to encourage those text nodes and visual nodes that serve a similar role in the visual SG^V and language SG^L to be closer, while for those not we hope to repel them from each other, so as to mitigate the scene inconsistency. We realize this via the current popular CL technique. We
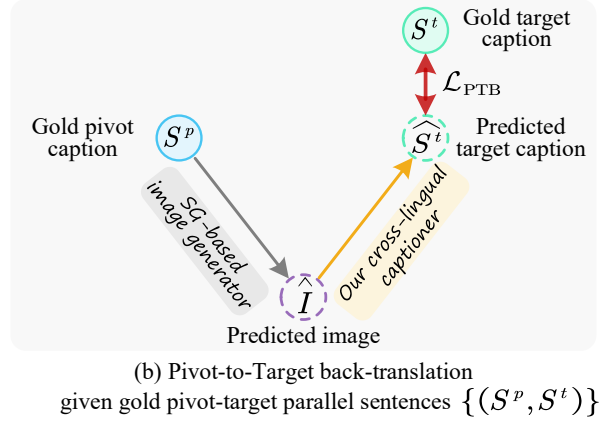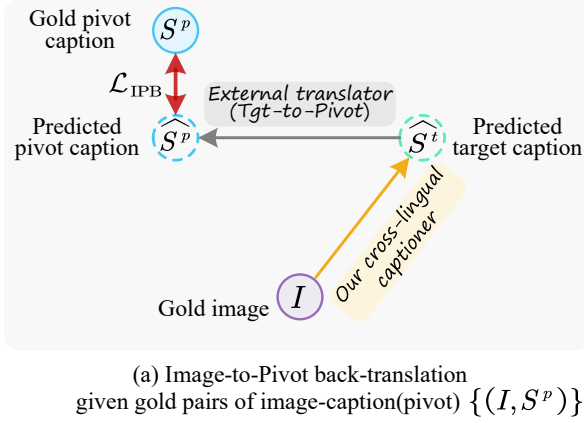
(a) Image-to-Pivot back-translation
given gold pairs of image-caption(pivot) $\{(I, S^p)\}$

(b) Pivot-to-Target back-translation
given gold pivot-target parallel sentences $\{(S^p, S^t)\}$

Figure 4: Illustrations of the cross-modal and cross-lingual back-translation.

first obtain the node representations of visual SG (h_i^V) and language SG (h_j^L) using one shared GCN encoder as in Eq. (1), based on the ground-truth $\{(I, S^p)\}$ data. We then measure the similarities between all pairs of nodes from two SGs:

$$s_{i,j}^m = \frac{(h_i^V)^T \cdot h_j^L}{||h_i^V|| \, ||h_j^L||}. \quad (7)$$

A pre-defined threshold ρ_m will decide the alignment confidence, i.e., pairs with $s_{i,j}^m > \rho_m$ is considered similar. Then we have:

$$\mathcal{L}_{\text{CMA}} = - \sum_{i \in \text{SG}^V, \, j^* \in \text{SG}^L} \log \frac{\exp(s_{i,j^*}^m / \tau_m)}{\mathcal{Z}}, \quad (8)$$

where $\tau_m > 0$ is an annealing factor. j^* represents a positive pair with i, i.e., $s_{i,j^*}^m > \rho_m$. $\mathcal{Z}$ is a normalization factor for probability calculation.

4.2 Cross-lingual Syntactic Structure Aligning

The idea is similar to the above one, while in the cross-lingual syntactic structure space. We use the shared SC GCN encoder to generate node representations r_i^P and r_j^T of pivot-/target-side SCs on the parallel sentences. CL loss is then put on the similarity score $s_{i,j}^l$ to carry out the unsupervised alignment learning, which we summarize as $\mathcal{L}_{\text{CLA}}$.

4.3 Cross-modal&lingual Back-translation

Drawing inspiration from unsupervised machine translation, we leverage the back-translation technique (Sennrich et al., 2016; Edunov et al., 2018) to achieve the two-step alignment over the overall framework. We present the cross-lingual cross-modal back-translation training, including the image-to-pivot back-translation and the pivot-to-target back-translation.

Image-to-Pivot Back-translation With gold image-caption pairs at hand, we can first obtain the target caption prediction $\hat{S}^t$ via our cross-lingual

captioner. We then translate the $\hat{S}^t$ into pseudo pivot caption $\hat{S}^p$ via an external translator $\mathcal{M}_{t \to p}$. This thus forms a path: $S^p\text{-}I \to \hat{S}^t \to \hat{S}^p$. And our framework can be updated via:

$$\mathcal{L}_{\text{IPB}} = \mathbb{E}[-\log p(\hat{S}^p | \mathcal{M}_{t \to p}(\mathcal{F}_{I \to S^t}(I)))]. \quad (9)$$

Pivot-to-Target Back-translation There is a similar story for the gold pivot-target parallel sentences: $S^t\text{-}S^p \to \hat{I} \to \hat{S}^t$. For $S^p \to \hat{I}$ we leverage an external SG-based image generator (Johnson et al., 2018; Zhao et al., 2022). The learning loss is:

$$\mathcal{L}_{\text{PTB}} = \mathbb{E}[-\log p(\hat{S}^t | \mathcal{F}_{I \to S^t}(\mathcal{M}_{S^p \to I}(S^p)))]. \quad (10)$$

★ **Remarks on Training** We take a warm-start strategy to ensure stable training of our framework. Initially we pre-train two parts separately via $\mathcal{L}_{\text{Cap}}$&$\mathcal{L}_{\text{Trans}}$ We then perform two structure-pivoting unsupervised alignment learning via $\mathcal{L}_{\text{CMA}}$&$\mathcal{L}_{\text{CLA}}$. Finally, we train the overall model via back-translation $\mathcal{L}_{\text{IPB}}$&$\mathcal{L}_{\text{PTB}}$. Once the system tends to converge, we put them all together for further overall fine-tuning:

$$\mathcal{L} = \mathcal{L}_{\text{Cap}} + \mathcal{L}_{\text{Trans}} + \mathcal{L}_{\text{CMA}} + \mathcal{L}_{\text{CLA}} + \mathcal{L}_{\text{IPB}} + \mathcal{L}_{\text{PTB}}. \quad (11)$$

Here for brevity, we omit the item weights. Appendix §A.4 gives more training details.

5 Experimental Setups

Datasets To align with existing work, we consider the transfer between English (En) and Chinese (Zh), and use the image caption datasets of MSCOCO (Lin et al., 2014), AIC-ICC (Wu et al., 2017) and COCO-CN (Li et al., 2019). We use the training set of a language as image-pivot pairs for the first part training, and test with the set of another language. For the second part training, we collect the paired En-Zh parallel sentences from existing MT data, including UM (Tian et al., 2014) and WMT19 (Barrault et al., 2019).

	Zh → En				En → Zh				Avg.
	BLEU	METEOR	ROUGE	CIDEr	BLEU	METEOR	ROUGE	CIDEr	
• *Translation-based methods*									
EarlyTranslation	48.3	15.2	27.2	18.7	43.6	20.3	30.3	14.2	27.2
LateTranslation	45.8	13.8	25.7	14.5	41.3	13.5	26.7	14.0	24.4
FG	46.3	12.5	25.3	15.4	43.0	19.7	29.7	15.7	25.9
SSR[†]	52.0	14.2	27.7	28.2	46.0	22.8	32.0	18.3	30.1
• *Pivoting-based methods*									
PivotAlign	52.1	17.5	28.3	27.0	47.5	23.7	32.3	19.7	31.1
UNISON	54.3	18.7	30.0	28.4	48.7	25.2	33.7	21.9	32.4
CROSS^{2}STRA (Ours)	**57.7**	**21.7**	**33.5**	**30.7**	**52.8**	**27.6**	**36.1**	**24.5**	**35.8**
w/o SG	55.8	19.1	31.2	28.0	48.6	25.8	33.9	21.6	33.1
w/o SC	56.1	20.0	32.1	28.9	50.4	26.6	35.4	23.3	34.1
w/o ResiConn	56.4	21.2	32.9	29.4	51.8	27.1	35.9	24.1	34.9

Table 1: Transfer results between MSCOCO (En) and AIC-ICC (Zh). The values of SSR[†] are copied from Song et al. (2019), while all the rest are from our implementations.

Baselines and Evaluations Our comparing systems include 1) the translation-based methods, including the *early translation* and *late translation* mentioned in the introduction, *FG* (Lan et al., 2017), *SSR* (Song et al., 2019), and 2) the pivoting-based methods, including *PivotAlign* (Gu et al., 2018b) and *UNISON* (Gao et al., 2022). Following baselines, we report the BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015) scores for model evaluation. Our results are computed with a model averaging over 10 latest checkpoints.

Implementations To obtain the visual SGs, we employ the FasterRCNN (Ren et al., 2015) as an object detector, and MOTIFS (Zellers et al., 2018) as a relation classifier and an attribute classifier. For language SGs, we first convert the sentences into dependency trees with a parser (Anderson et al., 2018), and then transform them into SGs based on certain rules (Schuster et al., 2015). We obtain the SC trees via the Berkeley Parser (Kitaev and Klein, 2018), trained on PTB (Marcus et al., 1993) for En texts and on CTB (Xue et al., 2005) for Zh texts. In our back-translation learning, we use the T5 (Raffel et al., 2020) as the target-to-pivot translator ($\mathcal{M}_{t \to p}$), and adopt the current SoTA SG-based image generator ($\mathcal{M}_{S^p \to I}$) (Zhao et al., 2022). Chinese sentences are segmented via Jieba[3]. We use Transformer to offer the underlying textual representations for GCN, and use FasterRCNN (Ren et al., 2015) for encoding visual feature representations. Our SG and SC GCNs and all other embeddings have the same dimension of 1,024. All

---

	Zh → En		En → Zh		Avg.
	B	R	B	R	
CROSS^{2}STRA	**57.7**	**33.5**	**52.8**	**36.1**	**45.0**
w/o L_{CMA}	54.4	29.7	50.1	34.9	42.3(-2.7)
w/o L_{CLA}	54.6	30.1	51.0	35.3	43.0(-2.0)
w/o L_{IPB}	53.8	31.1	50.5	35.1	43.1(-1.9)
w/o L_{PTB}	55.0	32.8	52.2	35.7	44.2(-0.8)
w/o L_{CMA}+L_{CLA}	51.8	27.7	47.5	33.7	40.8(-4.2)
w/o L_{IPB}+L_{PTB}	52.7	30.1	49.9	34.2	42.2(-2.8)

Table 2: Ablating different learning strategies. B: BLEU, R: ROUGE.

models are trained and evaluated with NVIDIA A100 Tensor Core GPUs.

6 Experimental Results and Analyses

Transfer between MSCOCO and AIC-ICC Table 1 presents the Zh→En and En→Zh transfer results. We first can observe that the *EarlyTranslation* is more effective than *LateTranslation*, as the former introduces lesser noises in training. Also, we see that among all the translation-based methods, *SSR* shows the best performance. Further, it is clear that the pivoting methods show overall better results than the translation ones. This is most possibly because the joint training in pivoting-based models relieves the under-alignment between the captioning and translation stages, reducing the noise introduction of the pipeline.

Looking into the pivoting-based models, *UNISON* exhibits the stronger capability of the transfer in both directions, owing to the integration of SG structure features, i.e., helping accurately capture the semantic relevances between vision and language. Most importantly, our CROSS^{2}STRA outperforms all the other baselines with significant

	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE	CIDEr	*Avg.*
• *Translation-based methods*								
EarlyTranslation[†]	60.4	40.7	26.8	17.3	24.0	43.6	52.7	37.9
LateTranslation[†]	58.9	38.0	23.5	14.3	23.5	40.2	47.3	35.1
SSR	65.2	43.5	27.3	17.7	25.4	45.9	53.8	39.8
• *Pivoting-based methods*								
PivotAlign	66.5	45.0	29.3	18.2	27.0	46.3	55.0	41.0
UNISON[*†]	63.4	43.2	29.5	17.9	24.5	45.1	53.5	39.5
UNISON	68.3	46.7	30.6	19.0	29.4	48.0	56.3	42.7
CROSS²STRA	**70.4**	**48.8**	**32.5**	**20.8**	**31.9**	**50.6**	**58.2**	**44.7**

Table 3: Cross-lingual transfer from MSCOCO (En) to COCO-CN (Zh). The values with † are copied from Song et al. (2019). UNISON*† is the raw version without using the paired image-caption(pivot) data for training.
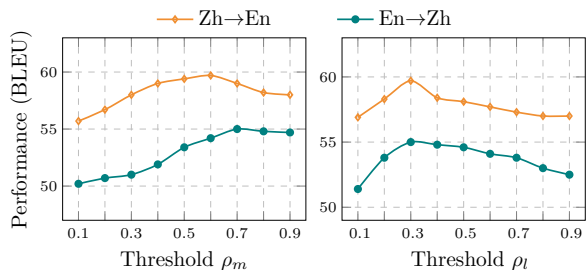

Figure 5: Influences of different threshold values in two structure-guided alignment learning.

margins on all metrics consistently. For example, we improve over *UNISON* by 3.4 (Zh→En) and 4.1 (En→Zh) BLEU scores respectively. We give credit to the integration of both the semantic SG and the syntactic SC structures, as well as the effective alignment learning strategies. The above observations show the efficacy of our system for cross-lingual captioning.

Influences of Learning Strategies In Table 2 we quantify the contribution of each learning objective via ablation. As seen, each learning strategy shows the impact to different extents. For example, the cross-modal semantic alignment gives greater influences than the cross-lingual syntactic alignment of the overall performances (i.e., 2.7 vs. 2.0). In contrast to the two structure-pivoting learning ($L_{CMA}+L_{CLA}$), we can find that the back-translation learning ($L_{IPB}+L_{PTB}$) shows slightly lower impacts. Particularly the pivot-to-target back-translation contributes limitedly, and we believe the quality of SG-to-image generator should bear the responsibility.

Threshold Study In Fig. 5 we study the influences of threshold values on the two alignment learning, by varying ρ_m and ρ_l. As seen, when ρ_m is 0.6 and 0.7 in two tasks respectively, the overall transfer results are the best, while ρ_l=0.3 helps give the best effects. Such a pattern distinction
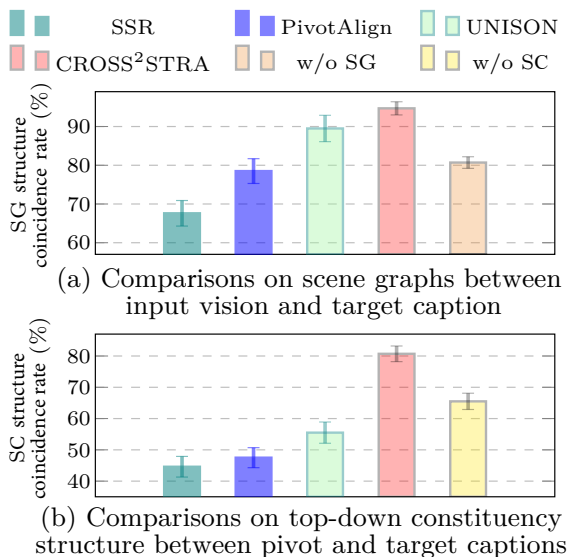

(a) Comparisons on scene graphs between input vision and target caption

(b) Comparisons on top-down constituency structure between pivot and target captions

Figure 6: The matchings of SG and SC structures.

between ρ_m and ρ_l implies that the SGs between vision and language have less discrepancy, while the SC structures between two languages come with non-negligible differences.

Transfer from MSCOCO to COCO-CN Table 3 further shows the transfer results from English MSCOCO to Chinese COCO-CN. The overall tendency is quite similar to the one in Table 1. We see that translation methods are inferior to the pivoting methods. Our CROSS²STRA model gives the best performances on all metrics, outperforming *UNISON* by an average 2.0(=44.7-42.7) score. This again verifies the efficacy of our proposed method.

Probing Cross-modal and Cross-lingual Structure Alignment We integrate the semantic scene structure and syntactic structures with the aim of better cross-modal and cross-lingual alignment in our two-stage pivoting transfer framework. Here we directly assess to what extent our methods improve the alignment. Fig. 6 shows the structure

| | | Gold | 绿油油的草地上蹲着一个穿着灰色寸衫面带微笑的小朋友
(A smiling child in grey shirt is squatting on the green grass) | | | Gold | 一名足球队员正在足球场上与另一名足球的运动员争夺足球
(A football player is competing for football with another football player on the football field) |

Figure 7: Qualitative results of cross-lingual captioning. The instances are randomly picked from AIC-ICC (Zh).

	Relevancy↑	Diversification↑	Fluency↑
FG	5.34	3.75	7.05
SSR	7.86	5.89	7.58
PivotAlign	8.04	6.57	7.46
UNISON	9.02	9.14	7.89
CROSS^{2}STRA	**9.70**‡	**9.53**‡	**9.22**‡
w/o SG	8.35	7.75	9.04
w/o SC	9.42	8.34	8.07
w/o L_{CMA}+L_{CLA}	7.80	7.24	8.15

Table 4: Human evaluations are rated on a Likert 10-scale. ‡ indicates significant better over the baselines (p<0.03).

Figure 8: Evaluating the language quality of captions.

coincidence rate between the input image SG and predicted target caption SG, and the SC structure coincidence rate between the pivot and target captions.[4] We see that with the integration of semantic scene modeling, both *UNISON* and our system exhibit prominent cross-modal alignment ability, i.e., with higher structural overlaps. The same observation can be found with respect to syntactic structure integration for enhancing cross-lingual alignment learning. Either without the leverage of SG or SC structure, the corresponding cross-modal or cross-lingual alignment effect is clearly weakened.

Human Evaluation We further try to quantify the improvements of the generated captions via human evaluation. In Table 4 we show the evaluation results based on MSCOCO (En) to AIC-ICC (Zh) transfer, on three dimensions: *relevancy*, *diversification* and *fluency*. We can see that our system shows significantly higher scores than baseline sys-
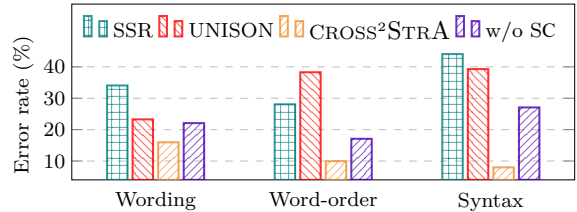
[4]Appendix §B.2 details the measuring method.

tems in terms of all three indicators. For those methods with SG structure features, the content relevancy and diversification of captions are much better. Yet only our method gives satisfied language fluency, due to the equipment of syntactic features. With further ablation studies we can further confirm the contributions of the SG and SC features.

Captioning Linguistic Quality Study We take a further step, investigating how exactly our model improves the linguistic quality of the target captions. Same to the human evaluation, we ask native speakers to measure the errors that occurred in the generated captions, in terms of *wording*, *word order* and *syntax correctness*. Fig. 8 presents the results of the transfer from MSCOCO (En) to AIC-ICC (Zh). We see that our model has committed the least errors, where the performances on syntax correctness are especially higher than baselines. Once without using the syntactic features, the error rates grow rapidly, which demonstrates the importance to integrate the syntactic structures.

Qualitative Result Finally, we empirically show some real prediction cases, so as to aid an intuitive

understanding of our method's strength. In Fig. 7 we provide four pieces of testing examples on the En→Zh transfer, which we compare with different baseline methods. As can be seen, the *SSR* model often tends to generate target-side captions with lower diversification, and meanwhile unsatisfactory content relevancy, and thus inaccurate image descriptions. On the contrary, the captions from *UNISON* are much better, i.e., better relevancy and diversification. We can give credit to the equipment of scene graph-based alignment learning. However, *UNISON* can fall short on language quality, i.e., problematic fluency. Since English and Chinese differ much in linguistic and grammar characteristics, without leveraging the syntactic structure features, it leads to inferior language quality. Luckily, our model can address all those issues, and generate captions with good relevancy, diversification, and fluency. This again proves the effectiveness of our proposed method.

7 Conclusion and Future Work

In this paper we investigate the incorporation of semantic scene graphs and syntactic constituency structure information for cross-lingual image captioning. The framework includes two phrases, semantic structure-guided image-to-pivot captioning and syntactic structure-guided pivot-to-target translating. We take the SG and SC structures as pivots, performing cross-modal semantic structure alignment and cross-lingual syntactic structure alignment learning. A cross-lingual&cross-modal back-translation training is further performed to align two phrases. On English↔Chinese transfer experiments, our model shows great superiority in terms of captioning relevancy and fluency.

Bridging the gaps between the cross-modal and cross-lingual transfer with external semantic and syntactic structures has shown great potential. Thus it is promising to extend the idea to other scenarios. Also, exploiting the external structures potentially will introduce noises, and thus a dynamical structure induction is favorable.

Limitations

In this work, we take the sufficient advantages of the external semantic and syntactic structure knowledge to improve our focused problem. But this could be a double-edged sword to use such features. Specifically, our paper has the following two potential limitations. First of all, our method closely relies on the availability of the resources of scene graph structures and syntax structures. While most of the languages come with these structure annotations to train good-performing structure parsers (for example, the syntax structure annotations of Penn TreeBank cover most of the existing languages), some minor languages may not have structure resources. That being said, our idea still works well even in the absence of the target-side structure annotations. With only the structure annotations at pivot-side (resource-rich) language (in this case, the cross-modal semantic&syntactic structure aligning learning are canceled), we can still achieve much better performances than those baselines without using the structural features. Besides, our method will be subject to the quality of the external structure parsers. When the parsed structures of scene graphs and syntax trees are with much noise, the helpfulness of our methods will be hurt. Fortunately, the existing external semantic and syntactic structure parsers have already achieved satisfactory performances, which can meet our demands.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation*, pages 1–61.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10575–10584.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5687–5698.

Hao Fei, Fei Li, Bobo Li, and Donghong Ji. 2021. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12794–12802.

Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2022. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475.

Jiahui Gao, Yi Zhou, Philip L. H. Yu, Shafiq R. Joty, and Jiuxiang Gu. 2022. UNISON: unpaired cross-lingual image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10654–10662.

Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. 2018a. Stack-captioning: Coarse-to-fine learning for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6837–6844.

Jiuxiang Gu, Shafiq R. Joty, Jianfei Cai, and Gang Wang. 2018b. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision*, pages 519–535.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237.

Paria Jamshid Lou, Yufei Wang, and Mark Johnson. 2019. Neural constituency parsing of speech transcripts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2756–2765.

Paria Jamshid Lou, Yufei Wang, and Mark Johnson. 2020. Improving disfluency detection by self-training a self-attentive model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3754–3763.

Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. 2021. Improving image captioning by leveraging intra- and inter-layer global representation in transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1655–1663.

Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2407–2415.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2676–2686.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In *Proceedings of the ACM International Conference on Multimedia*, pages 1549–1557.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360.

Zuchao Li, Masao Utiyama, Eiichiro Sumita, and Hai Zhao. 2021. Unsupervised neural machine translation with universal grammar. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3249–3264.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *Proceedings of the International Conference on Learning Representations*.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Proceedings of the Computational Linguistics and Intelligent Text Processing*, pages 3–16.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 91–99.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 70–80.

Lane Schwartz, Chris Callison-Burch, William Schuler, and Stephen Wu. 2011. Incremental syntactic language models for phrase-based translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 620–631.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Yuqing Song, Shizhe Chen, Yida Zhao, and Qin Jin. 2019. Unpaired cross-lingual image caption generation with self-supervised rewards. In *Proceedings of the ACM International Conference on Multimedia*, pages 784–792.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. 2014. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1837–1842.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

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

Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. 2018. Scene graph parsing as dependency parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 397–407.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. 2017. AI challenger : A large-scale dataset for going deeper in image understanding. *CoRR*, abs/1711.06475.

Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. 2021. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963.

Hongfei Xu, Josef van Genabith, Deyi Xiong, Qiuhui Liu, and Jingyi Zhang. 2020. Learning source phrase representations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 386–396.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 5065–5075.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840.

Jiajun Zhang and Chengqing Zong. 2013. Learning a phrase-based translation model from monolingual data with application to domain adaptation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1425–1434.

Yue Zhang and Zhenghua Li. 2022. Csyngec: Incorporating constituent-based syntax for grammatical error correction with a tailored gec-oriented parser. *CoRR*, abs/2211.08158.

Xin Zhao, Lei Wu, Xu Chen, and Bin Gong. 2022. High-quality image generation from scene graphs with transformer. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6.

A Model Details

A.1 Specification of Scene Graph and Syntax Constituency Structures

In Fig. 9 and Fig. 10 we illustrate the complete structures of the syntactic constituency tree and scene graphs, respectively. We note that the scene graph is a dependency-like structure, describing the node-node inter-relation in an 'is-a' paradigm. And the syntactic constituency tree is a compositional structure, depicting how words constitute phrases and then organize them into whole sentences.

A.2 Pivot-to-target Translation

In Eq. (4) we use a Transformer decoder to predict the target caption $\hat{S}^t$. A cross-attention mechanism is first used to fuse the prior SG feature representations $\boldsymbol{h}$ and the SC features $\boldsymbol{r}$. Specifically,

$$\boldsymbol{e} = \text{Softmax}(\frac{\boldsymbol{r} \oplus \boldsymbol{h}}{\sqrt{d}}) \cdot \boldsymbol{r} \,,$$

where d is a scaling factor. Then, the Transformer performs decoding over $\{\boldsymbol{e}\}$:

$$\hat{S}^t = \text{Trm}^C(\{\boldsymbol{e}\}) \,.$$

A.3 Specification on Contrastive Learning

Cross-modal Semantic Structure Aligning In Eq. (8) we define the contrastive learning objective of cross-modal semantic structure aligning, here we unfold the equation:

$$\mathcal{L}_{\text{CMA}} = - \sum_{i \in \text{SG}^V, \, j^* \in \text{SG}^L} \log \frac{\exp(s_{i,j^*}^m / \tau_m)}{\mathcal{Z}} \,,$$
$$\mathcal{Z} = \sum_{i \in \text{SG}^V, \, k \in \text{SG}^L, \, k \neq j^*} \exp(s_{i,k} / \tau_m) \,,$$

where $\tau_m > 0$ is an annealing factor. j^* represents a positive pair with i, i.e., $s_{i,j^*}^m > \rho_m$.

**Cross-lingual Syntactic Structure Aligning** We detail the cross-lingual syntactic structure aligning learning objective here:

$$\mathcal{L}_{\text{CMA}} = - \sum_{i \in \text{SC}^P, \, j^* \in \text{SC}^T} \log \frac{\exp(s_{i,j^*}^l / \tau_l)}{\mathcal{Z}} \,,$$
$$\mathcal{Z} = \sum_{i \in \text{SC}^P, \, k \in \text{SC}^T, \, k \neq j^*} \exp(s_{i,k} / \tau_l) \,,$$

where $\tau_l > 0$ is an annealing factor. j^* represents a positive pair with i, i.e., $s_{i,j^*}^m > \rho_m$.

A.4 Specifying Overall Training Processing

The training of our framework is based on the warm-up strategy, including four stages.
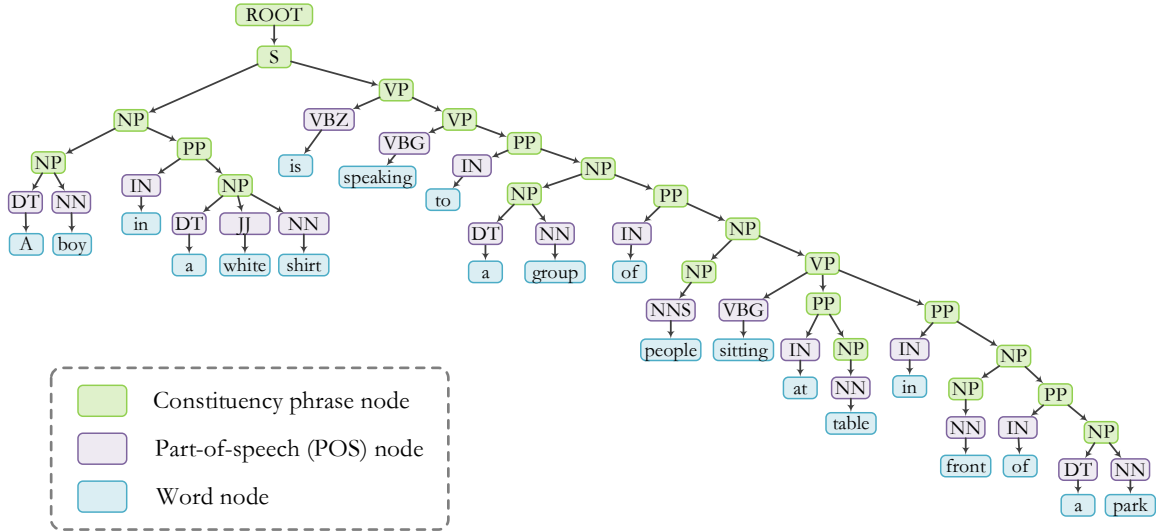
Input Sentence: A boy in a white shirt is speaking to a group of people sitting at table in front of a park

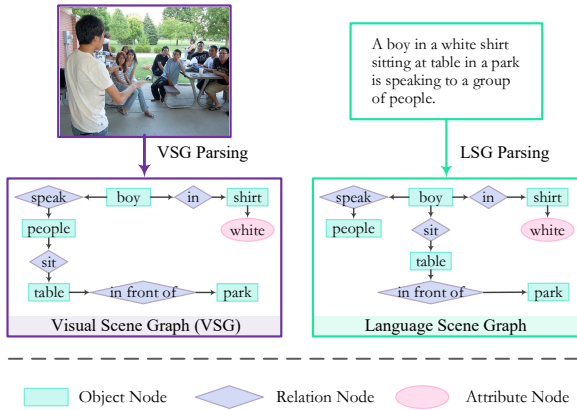Figure 9: Illustration of the full constituency syntax structure.

Figure 10: Illustration of the full scene graph structures. The object nodes are the real entities of visual proposals or textual words/phrases. The attribute nodes and relation nodes are the descriptive words, which, in VSG are automatically generated, and in LSG are extracted from sentences.

At the first stage, we use the paired image-caption data $\{(I, S^p)\}$ at the pivot language side (as well as the VSG structure features) to train the captioning part of our model; and use the parallel sentences $\{(S^p, S^t)\}$ (as well as the pivot-side syntax tree features) to train the translation part of our model, both of two training is supervised.

At the second stage, we perform two structure-pivoting unsupervised alignment learning, by using the image-caption data $\{(I, S^p)\}$, parallel sentences $\{(S^p, S^t)\}$, and the two structure resource.

At the third stage, we perform the cross-modal and cross-lingual back-translation learning. This is a whole-framework-level training, aiming to suffi-

ciently align the captioning and translation parts.

At the fourth stage, the system tends to converge, and we put all the above learning objects together for further overall fine-tuning:

$$\mathcal{L} = \lambda_{\text{Cap}}\mathcal{L}_{\text{Cap}} + \lambda_{\text{Trans}}\mathcal{L}_{\text{Trans}}$$
$$+ \lambda_{\text{CMA}}\mathcal{L}_{\text{CMA}} + \lambda_{\text{CLA}}\mathcal{L}_{\text{CLA}}$$
$$+ \lambda_{\text{IPB}}\mathcal{L}_{\text{IPB}} + \lambda_{\text{PTB}}\mathcal{L}_{\text{PTB}} .$$

Here λ_* are the loss weights that dynamically change by linearly learning scheduler (Huang et al., 2020). The initial weights are given as: λ_{Cap}=1, λ_{Trans}=1, λ_{CMA}=0.7, λ_{CLA}=0.7, λ_{VCB}=0.3, λ_{CPB}=0.3. λ_{Cap} and λ_{Trans} will be linearly decreased from 1 to 0.7 along the training, λ_{CMA} and λ_{REC} are kept unchanged, while λ_{VCB} and λ_{CPB} will be decreased from 0.3 to 0.7.

B Extended Experiment Setups

B.1 Dataset Description

We use three image captioning datasets $\{(I, S^p)\}$: MSCOCO, AIC-ICC and COCO-CN. All the data split follows the same practice as in prior cross-lingual image captioning works (Wu et al., 2017; Song et al., 2019). The MSCOCO dataset is annotated in English, which consists of 123,287 images and 5 manually labeled English captions for each image. We utilize 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. The AIC-ICC dataset contains 238,354 images and 5 manually annotated Chinese captions for each image. There are 208,354 and 30,000 images in the official training and validation set.

Dataset	Lang.	Split	#Image	#Caption
MSCOCO	En	Total	123,287	616,435
		Train	113,287	566,435
		Develop	5,000	25,000
		Test	5,000	25,000
AIC-ICC	Zh	Total	238,354	1,191,770
		Train	208,354	1,041,770
		Develop	25,000	125,000
		Test	5,000	25,000
COCO-CN	Zh	Total	20,342	27,218
		Train	18,342	25,218
		Develop	1,000	1,000
		Test	1,000	1,000

Table 5: Statistics of image captioning datasets.

Since the annotations of the testing set are unavailable in the AIC-ICC dataset, we randomly sample 5,000 images from its validation set as our testing set. The COCO-CN dataset contains 20,342 images and 27,218 caption texts in Chinese. We use 18,342 images for training, 1,000 for development, and 1,000 for testing. Table 5 gives the detailed statistics of the image captioning data.

For the translation data $\{(S^p, S^t)\}$, we collect about 1M of raw paired En-Zh parallel sentences from the UM (Tian et al., 2014) and WMT19 (Barrault et al., 2019) machine translation corpus. We filter the sentences in MT datasets according to an existing caption-style dictionary and resulting in a total of 400,000 parallel sentences. For the translation training, we use 390,000 sentence pairs for training, 5,000 sentence pairs for validation, and 5,000 pairs for testing.

B.2 Specification on Structure Coincidence Probing

In Fig. 6 we assess the ability of our model on the cross-modal and cross-lingual structure alignment, by measuring the structure coincidence between the gold one and the one learned by our model. Here we detail the evaluation setup.

For the semantic scene structures, we evaluate the coincidence between the input images' SGs and the SGs of predicted target-side captions. These SG structures are parsed by the same methods introduced above. We then make statistics of the overlapped node pairs between the two SGs as the coincidence rate β^G.

$$\beta^G = \frac{\mathrm{SG}^V \cap \mathrm{SG}^L}{\mathrm{SG}^V \cup \mathrm{SG}^L},$$

where SG^V and SG^L denote any word-pair structure of visual SG and target language SG, respectively.

For the syntax structures, we evaluate the co-

incidence rate of the constituency tree structures between the intermediate pivot captions and the final predicted target-side captions. (Because the input images come without the syntax trees.) The SC structures of two languages are parsed using the parsers introduced above. We note that the divergences of syntax between two languages can be much larger, compared with the divergences of semantic scene structures. Different from the measurement for SG structure to traverse the whole graph equally, we measure the SC structure coincidence rate β^C in a top-down manner. Specifically, we traverse the constituency trees in a top-down order, and those matched phrasal nodes at a higher level (lower traversing depth from the root node) will have higher scores than those at a lower level.

$$\beta^C = \frac{(\mathrm{SC}^P \cap \mathrm{SC}^T) \cdot d}{\mathrm{SC}^P \cup \mathrm{SC}^T},$$

where SC^P and SC^T denote the phrasal constituent structures of the pivot and target language, respectively. d is a weight, which is defined as the reciprocal of a top-down traversing depth.

B.3 Specifications of Human Evaluation Standards

Table 4 shows the human evaluation results. Specifically, we design a Likert 10-scale to measure the relevancy, diversification, and fluency of the generated target-side captions. The 10-scale metrics are defined as: 1-Can't be worse, 2-Terrible, 3-Poor, 4-Little poor, 5-Average, 6-Better than average, 7-Adequate, 8-Good, 9-Very good, 10-Excellent. We ask ten native Chinese speakers to score the results. And for each result, we use the averaged scores.

In Fig. 8 we also measure the language quality of captions in terms of *wording*, *word order*, and *syntax correctness*. We ask the same ten native Chinese speakers to score the error degree of these metrics, each of which is defined as:

- **Wording**: Is the choice of words in the captions suitable and precise to describe the input images?

- **Word order**: Are the words, phrases, and components organized correctly and properly in captioning sentences?

- **Syntax correctness**: Are there syntactic errors in the caption texts? such as omitting or repeating words, mixing up verb tenses or verb conjugations, missing prepositions, etc.

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

5&6

☑ B1. Did you cite the creators of artifacts you used?
5&6

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix B

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix B

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Appendix B

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix B

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix B

C ☑ Did you run computational experiments?

Appendix B

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix B

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Appendix B

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix B

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Appendix B

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.