Scene Graph-driven Structured Vision-Language Learning

场景图驱动的结构化视觉语言跨模态学习

NExT++ Research Center, NUS

Hao Fei, Research Fellow

Aug. 23rd 2023

https://haofei.vip/

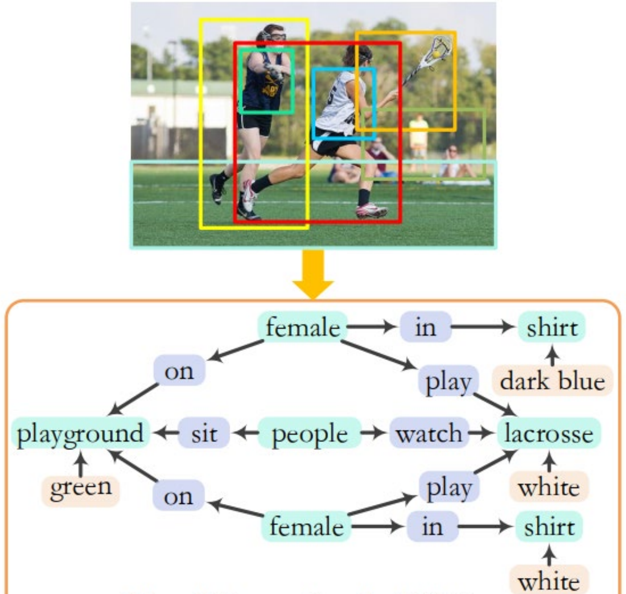
CONTENT

CONTENT

Vision&Language Scene Graphs

Scene Graph Representation

- Visual Scene Graph (VSG)
  - Representing visual content into semantic structured representation:
    - **Object Nodes:**
      Visually-seen entity objects
    - **Relation Nodes:**
      describing the semantic relations between objects
    - **Attribute Nodes**
      depicting the objects



Visual Scene Graph (VSG)

[1] *Justin Johnson, etc, and Li Fei-Fei. Image retrieval using scene graphs. CVPR. 2015.*

Vision&Language Scene Graphs

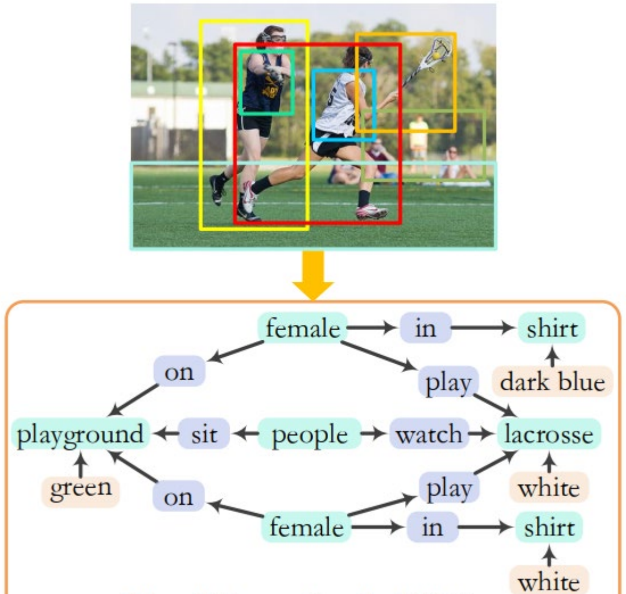
Scene Graph Representation

- VSG Parsing
  - Object detection

    e.g., FasterRCNN
  - Relation classification

    e.g., MOTIFS
  - Attribute classification

    e.g., MOTIFS

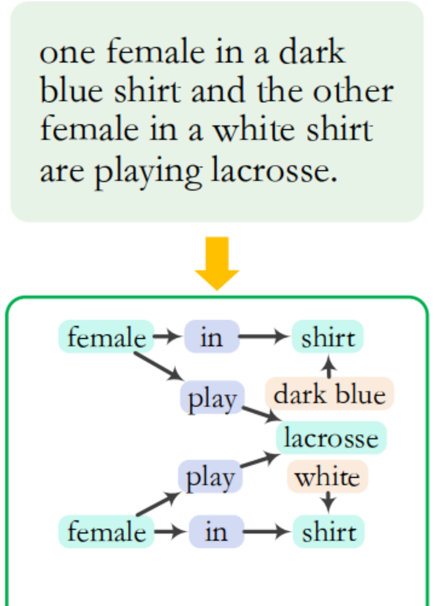


Visual Scene Graph (VSG)

Vision&Language Scene Graphs

Scene Graph Representation

- Language Scene Graph (LSG)
  - Representing textual inputs into semantic structured representation:
    - **Object Nodes:**
      entity tokens
    - **Relation Nodes:**
      verb/prep describing the semantic relations between objects
    - **Attribute Nodes**
      token/terms depicting the objects



Language Scene Graph (LSG)

[1] Yu-Siang Wang, etc. Scene graph parsing as dependency parsing. NAACL. 2018.

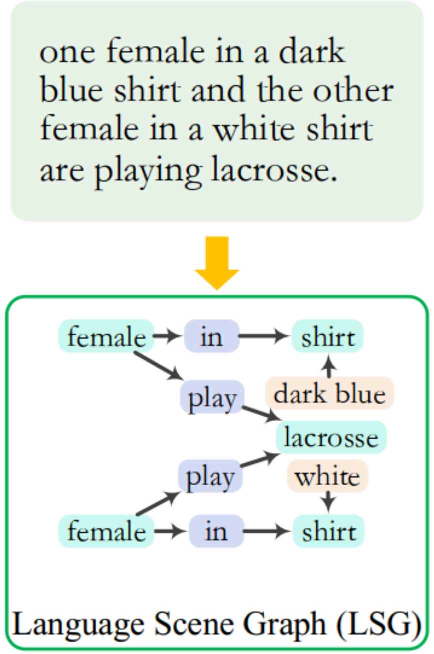
Vision&Language Scene Graphs

Scene Graph Representation

- LSG Parsing
  - Dependency Parsing

    e.g., Stanford Parser
  - Rule-based Conversion

    e.g., nsubj->object, adj->attribute

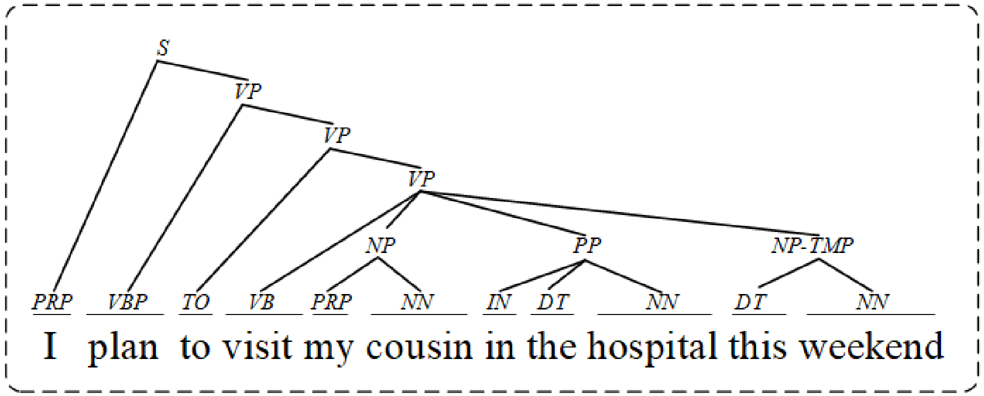


Scene Graph Representation

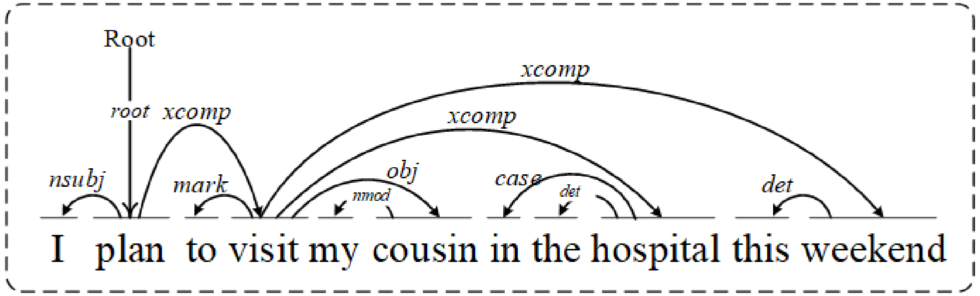
- Language Scene Graph (LSG)

There are so many structured representations of languages

- Syntactic-level structure



Constituency tree



Dependency tree

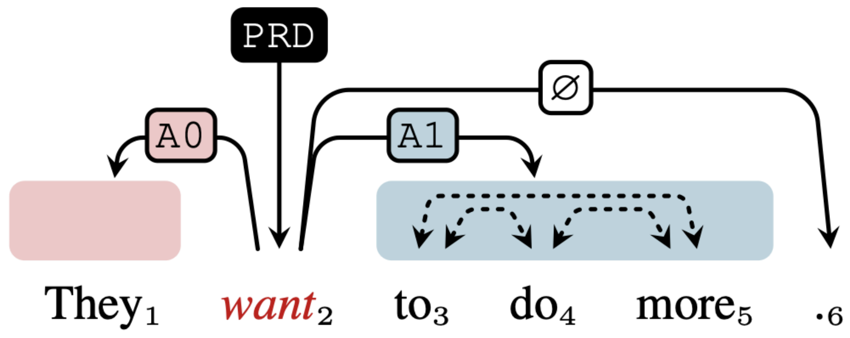
Vision&Language Scene Graphs

Scene Graph Representation

- Language Scene Graph (LSG)

There are so many structured representations of languages

- Semantic frame structure



Vision&Language Scene Graphs

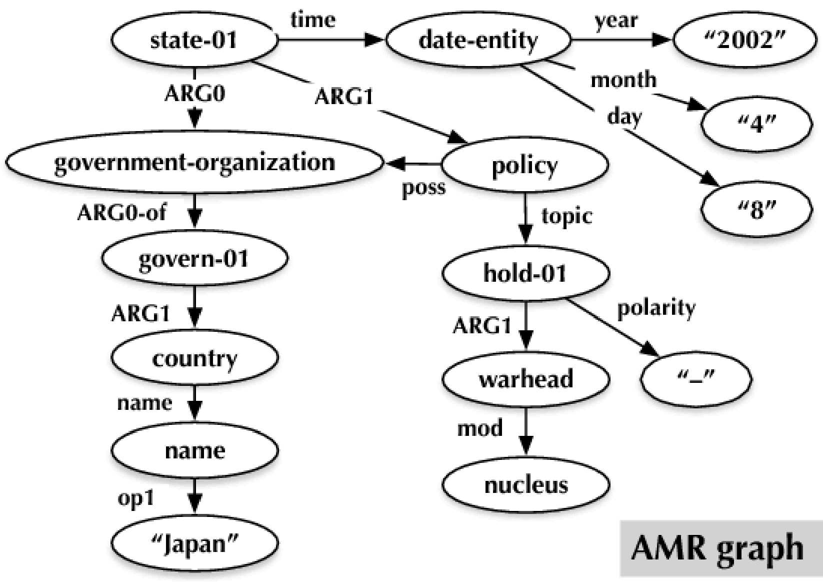
Scene Graph Representation

- Language Scene Graph (LSG)

There are so many structured representations of languages

- Semantic graph structure

The Japanese Government stated on April 8, 2002 its policy of holding no nuclear warheads.



AMR graph

```
(s / state-01
   :ARG0 (g / government-organization
      :ARG0-of (g2 / govern-01
         :ARG1 (c / country
            :name (n2 / name :op1 "Japan"))))
   :ARG1 (p / policy
      :poss g
      :topic (h / hold-01 :polarity -
         :ARG1 (w / warhead
            :mod (n / nucleus))))
   :time (d / date-entity :year 2002 :month 4 :day 8))
```

PENMAN format

Vision&Language Scene Graphs

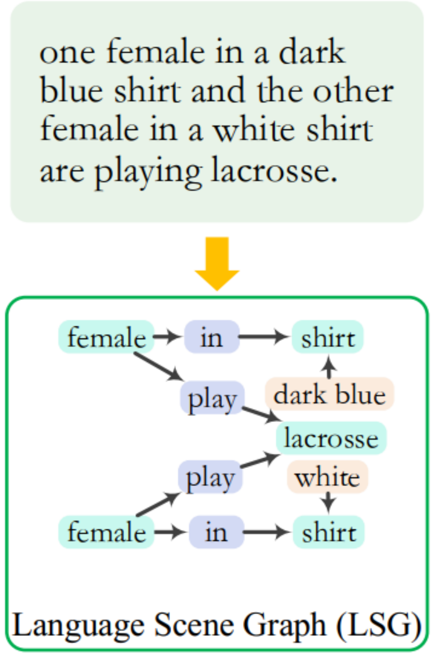
Scene Graph Representation

- Language Scene Graph (LSG)
  - Representing textual inputs into semantic structured representation:
    - **Object Nodes:**

      entity tokens
    - **Relation Nodes:**

      verb/prep describing the semantic relations between objects
    - **Attribute Nodes**

      token/terms depicting the objects



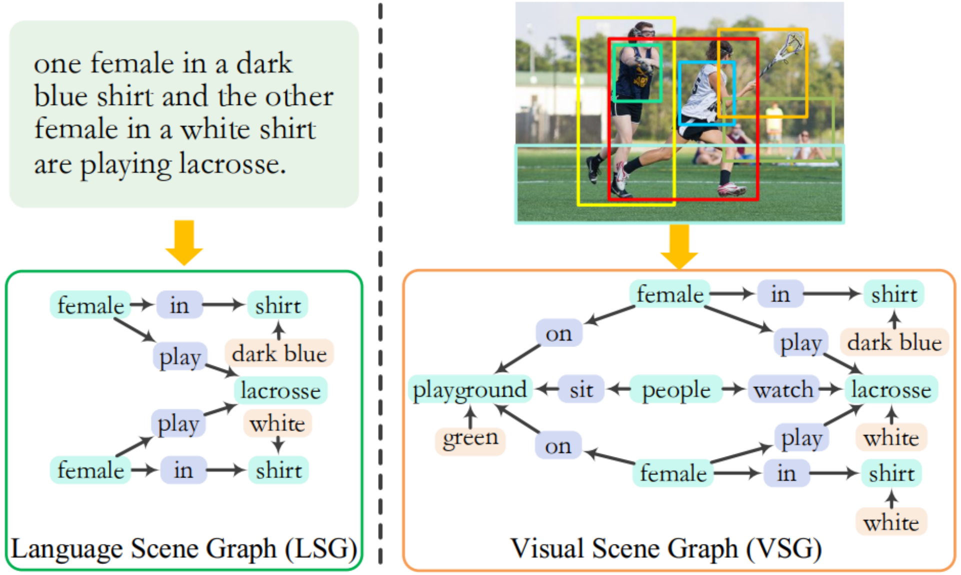
Vision&Language Scene Graphs

Scene Graph Representation

- Language Scene Graph (LSG)
  - Representing visual and textual inputs with **VSG** and **LSG**

- *The intrinsic gap between Vision and Language*

- *Unifying the Vision and Language with a unified representation format: **SG***



Vision&Language Scene Graphs

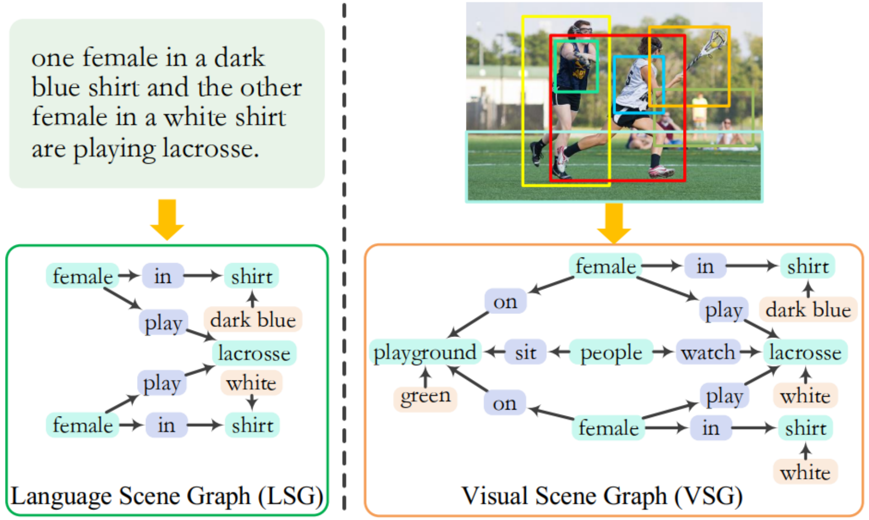
Scene Graph Representation

➢ Why do SG features help improve vision-language learning?

1. Improving cross-modal alignment:
   more fine-grained vision-text matching

2. Enhancing multimodal fusion:
   semantic-level feature learning

3. More controllable end-task prediction:
   highly structured modal representation



SG-based Cross-lingual Image Captioning

Application I:

CROSS²STRA: Unpaired Cross-lingual Image Captioning with Cross-lingual Cross-modal Structure-pivoted Alignment

[1] Shengqiong Wu, Hao Fei, Wei Ji, Tat-Seng Chua. Cross2StrA: Unpaired Cross-lingual Image Captioning with Cross-lingual Cross-modal Structure-pivoted Alignment. ACL. 2023.

SG-based Cross-lingual Image Captioning

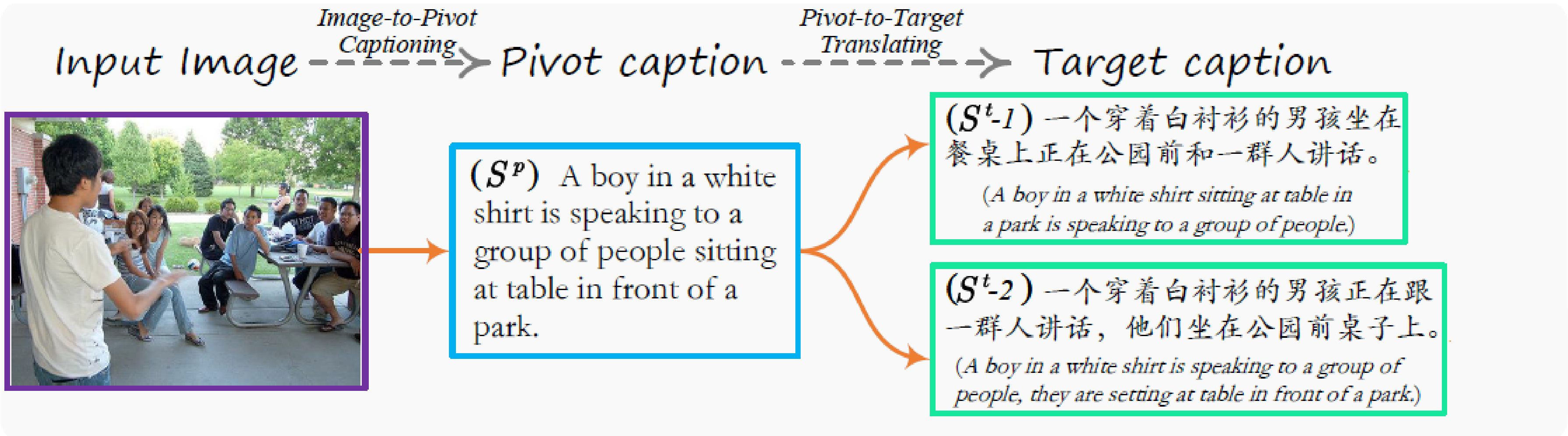
Motivation

➢ **Cross-lingual Image captioning**

- How to develop the image captioner in other languages, i.e., resource-scare language?

✓ The translation-based method

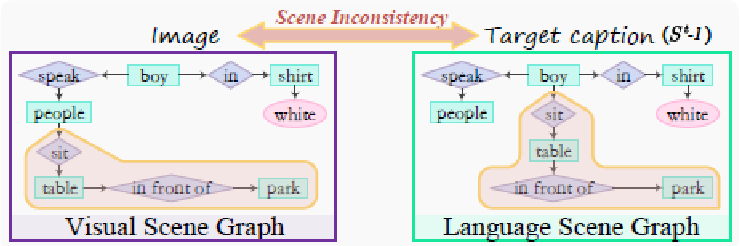
✓ The pivoting-based method



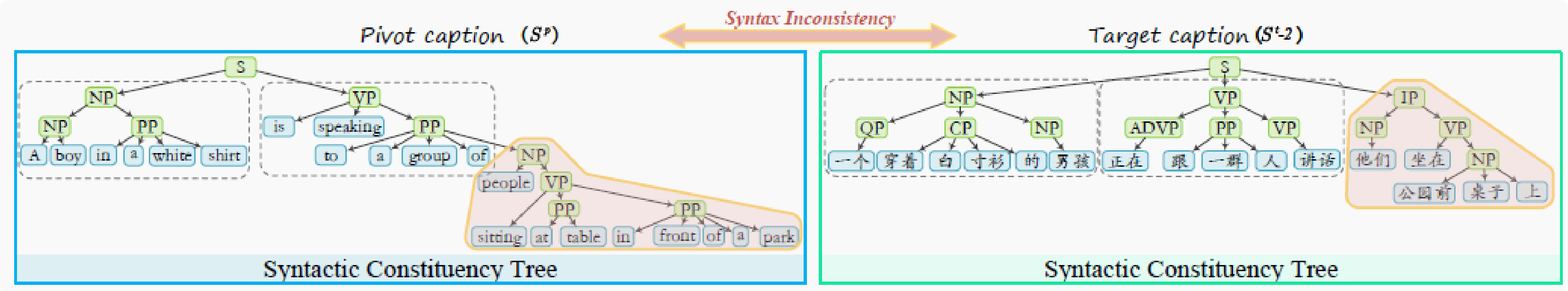
SG-based Cross-lingual Image Captioning

Motivation

- *irrelevancy*
- *disfluency*



(b) Relevancy issue due to inconsistency of semantic scene



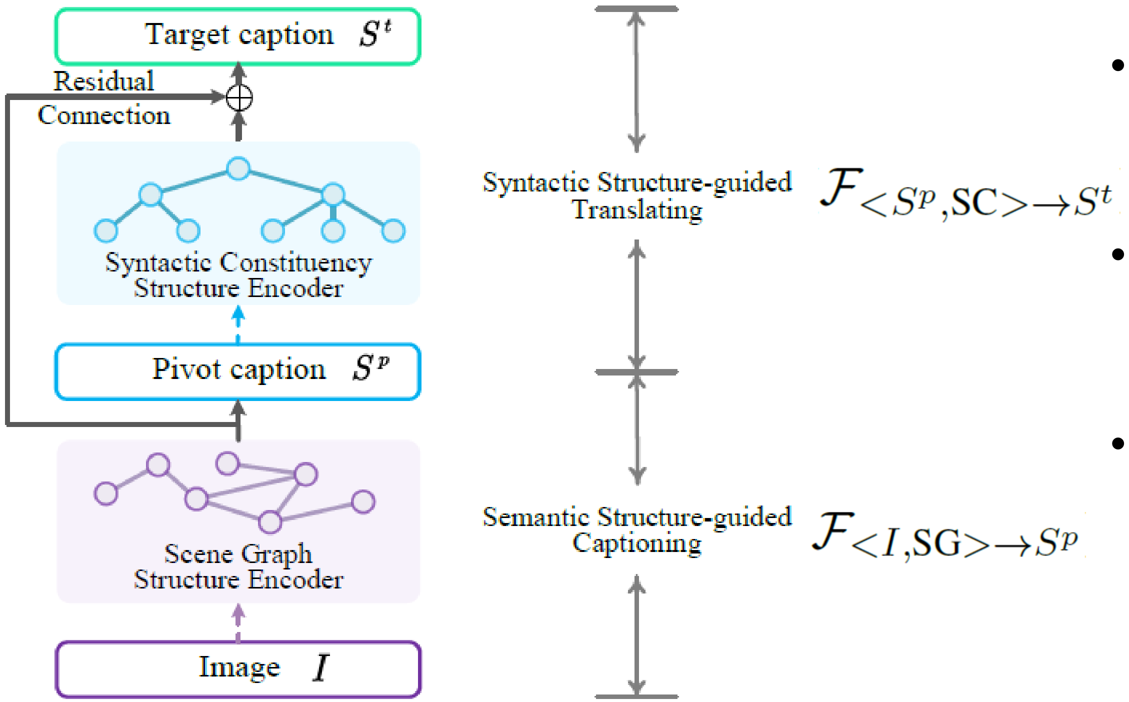
(c) Inconsistency of syntax structures between pivot-target languages causes disfluent translation

modeling the vision-language semantic alignment

modeling the pivot-target syntax alignment

SG-based Cross-lingual Image Captioning

Method

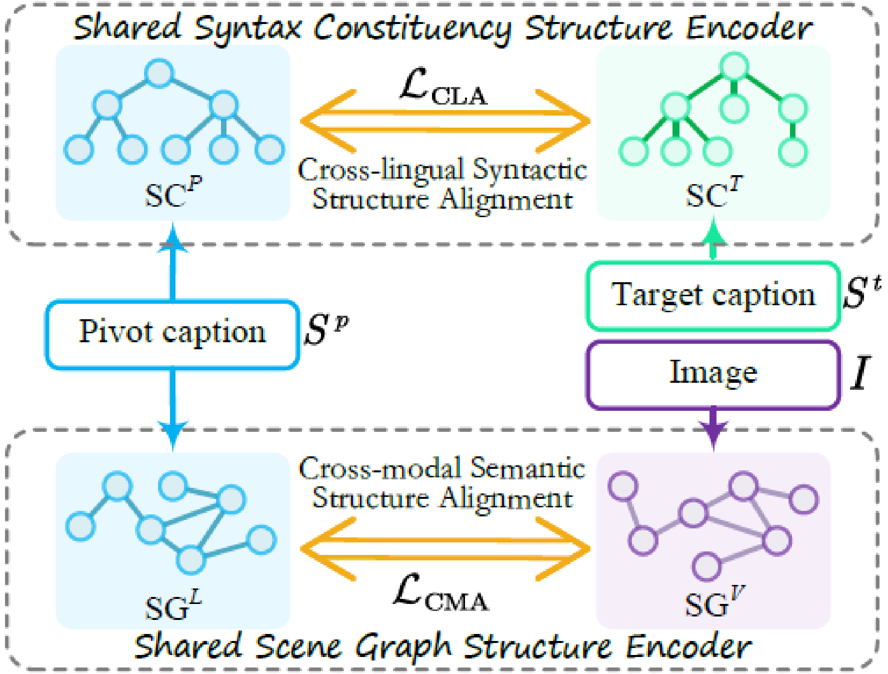


- A novel syntactic and semantic structure-guided model for cross-lingual image captioning
- For image-to-pivot captioning, we consider leveraging the **scene graphs (SG)** for better **image-text alignment**
- For the pivot-to-target translating, we make use of the **syntactic constituency (SC)** tree structures for better **pivot-target language alignment**.

SG-based Cross-lingual Image Captioning

Method

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



- Cross-modal Semantic Structure Aligning

  To encourage those text nodes and visual nodes that serve a similar role in the visual SG and language SG

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

similar pairs

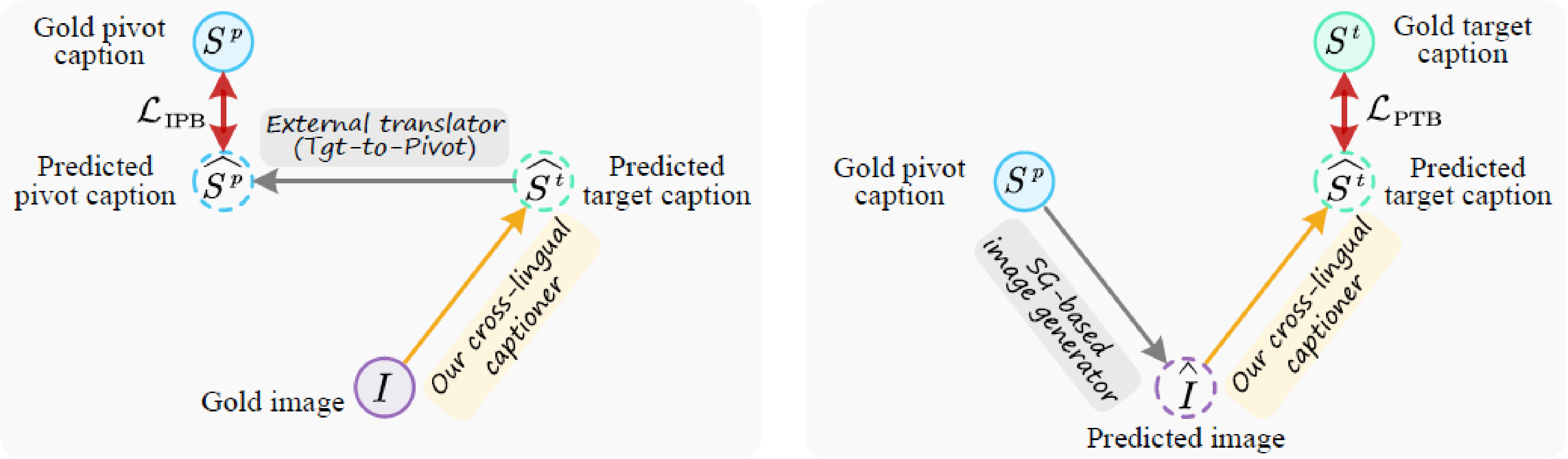
- Cross-lingual Syntactic Structure Aligning

$$\mathcal{L}_{\text{CLA}} = -\sum_{i \in \text{SC}^T, j^* \in \text{SC}^P} \log \frac{\exp(s^l_{i,j^*}/\tau_l)}{\mathcal{Z}}$$

Method

- Cross-modal&lingual Back-translation

To achieve the two-step alignment over the overall framework.



(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 $\{(S^p, S^t)\}$

$$S^p\text{-}I \rightarrow \hat{S}^t \rightarrow \hat{S}^p \qquad S^t\text{-}S^p \rightarrow \hat{I} \rightarrow \hat{S}^t$$

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

SG-based Cross-lingual Image Captioning

Experiment

➢ Transfer between MSCOCO and AIC-ICC

	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.

- *Pivoting* methods show overall better results than the *translation* ones
- CROSS^{2}STRA outperforms all the other baselines with significantly

Experiment

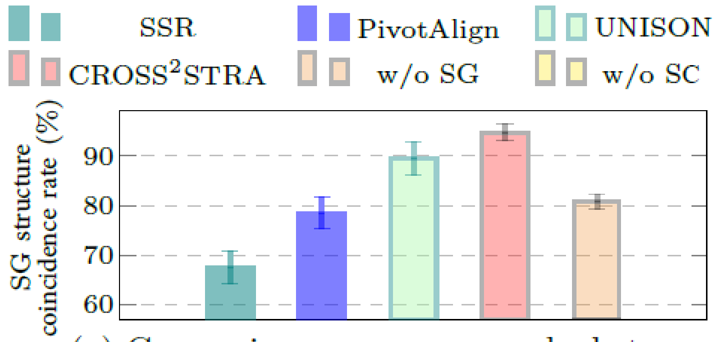
➢ Human Evaluation

	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
CROSS²STRA	**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

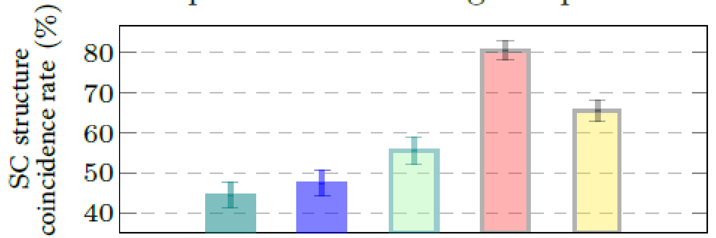
Table 4: Human evaluations are rated on a Likert 10-scale. ‡ indicates significant better over the baselines (p<0.03).

- Our system shows significantly higher scores than baseline systems in terms of all three indicators.
- With SG and SC structure features, the content relevancy and diversification of captions are much better

➢ Probing Cross-modal and Cross-lingual Structure Alignment



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



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

- Our system exhibit prominent structure alignment ability

SG-based Cross-lingual Image Captioning

Experiment

➢ Qualitative Result

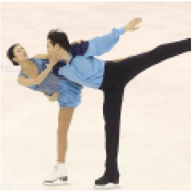
Image	Method	Caption
	Gold	绿油油的草地上蹲着一个穿着灰色寸衫面带微笑的小朋友 (A smiling child in grey shirt is squatting on the green grass)
	SSR	草丛上有一个小女孩坐在地上 (Sitting on the grass there is a little girl sitting on the ground)
	UNISON	坐在绿色草地上的小孩穿着灰色上衣 (Sitting on the green field is a kid wearing a gray coat)
	CROSS^{2}STRA	碧绿的草坪上蹲着一个满面笑容身穿灰色短袖的小孩 (A smiling kid in grey T-shirt is squatting on the green field)
	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)
	SSR	一个人在踢一个足球和一个红色衣服的男人 (A man is playing a football and a man in red)
	UNISON	在绿色的球场上一个身着白色衣服与一个红色衣服的男人踢白色的足球 (On the green football field, a man in white and a man in red play a white football)
	CROSS^{2}STRA	一位穿着白色球衣与另一外穿着红色球衣的运动员在绿色足球场上争夺足球 (A player wearing a white jersey and another player wearing a red jersey are competing for football on the green football field.)
	Gold	有一位拿着球拍的男运动员在球场上打网球 (There is a male player with a racket is playing tennis on the court)
	SSR	一个男人拿着网球拍 (A man is holding a tennis racket)
	UNISON	一位男运动员挥舞着网球拍在网球场 (A male athlete is waving a tennis racket on the tennis court)
	CROSS^{2}STRA	有一位身着白色衣裤的男性运动员拿着球拍在蓝色球场上打网球 (There is a male athlete in white clothes with a racket playing tennis on the blue court)
	Gold	滑冰场上，一名穿着黑色裤子的男士和一名穿着裙子的女人一起进行花样滑冰 (On the skating rink, a man in black trousers and a woman in skirt are doing figure skating together)
	SSR	穿着滑冰鞋身着演出服装的男人与女人在滑冰 (Men and women in skates and costumes are skating)
	UNISON	穿着蓝色衣服的男人和穿裙子的女人在滑冰场上进行花样滑冰 (The man in blue and the woman in skirt are figure skating on the skating rink)
	CROSS^{2}STRA	在滑冰场上有一名身着蓝色寸衫与黑色裤子的男士和一名穿着蓝色裙子的女士共同表演花样滑冰 (On the skating rink, a man in a blue shirt and black pants and a woman in a blue skirt together perform figure skating)

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

- With SG structure features, the content relevancy and diversification of captions are much better
- Our system generate captions with good relevancy, diversification, and fluency

SG-based Multimodal Relation Extraction

Application II:

***Information Screening whilst Exploiting!* Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling**

[1] Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, Tat-Seng Chua. Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling. ACL. 2023.

SG-based Multimodal Relation Extraction

Motivation

➢ Relation Extraction (RE)

Textual RE

member_of

Under *Cook*'s leadership, *Apple* has increased its donations to charity.
person company

Multimodal RE

- **Input Text:**

JFK and *Obama* at *Harvard* @Harvard
person person organization

- **Input Image:**

- **Output Relations:**
  (Graduated at, JFK, Harvard)
  (Graduated at, Obama, Harvard)
  (Alumni, JFK, Obama)

Supporting Visual Evidence: Bachelor Cap, Gown, Book

Motivation

- **Problem 1:** Internal-information over-utilization
  - *ONLY parts of the texts are useful to the relation inference*
  - *33.8% of tweets had textual content that was **not reflected in the images**, and the images did **not add additional content***

✓ A **fine-grained information pruning** over two modalities is needed

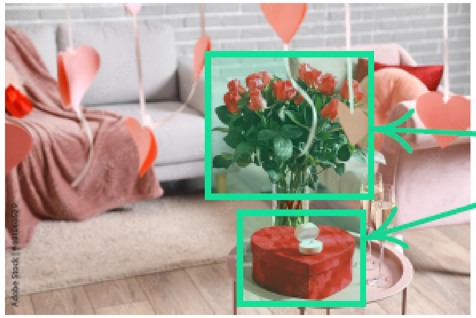
Last exam turned in. No more juggling work + school + family + hobbies. Maybe now they'll finally give me a BSc

Example #1

couple

Input Text: *Congratulations to Angela and Mark Salmons, a new life ahead is waiting!*

Input Image:



Useful feature for relation reasoning

Motivation

- **Problem 2:** External-information under-exploitation
  - *Short in text lengths and low-relevant images*

Information deficiency

✓ Additional **semantic supplementary information** is needed.

SG-based Multimodal Relation Extraction

Motivation

✓ A **fine-grained information pruning** over two muti-modalities is needed

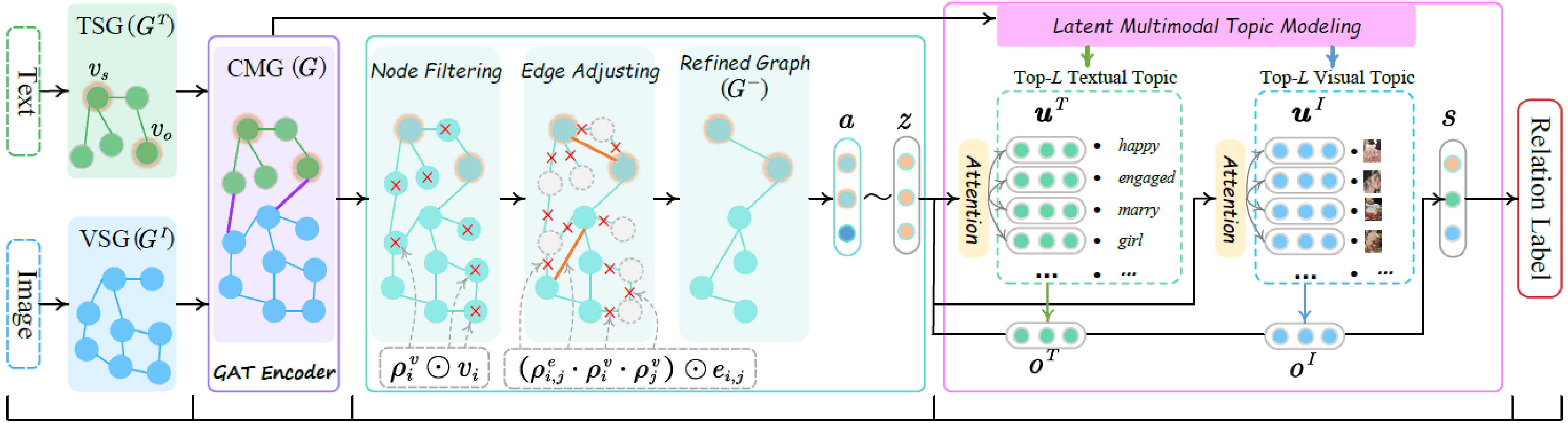
GIB-guided Feature Refinement

✓ Additional **semantic supplementary information** is needed.

Multimodal Topic Integration

SG-based Multimodal Relation Extraction

Method



Scene Graph Generation

Cross-modal Graph Construction

GIB-guided Feature Refinement

Multimodal Topic Integration

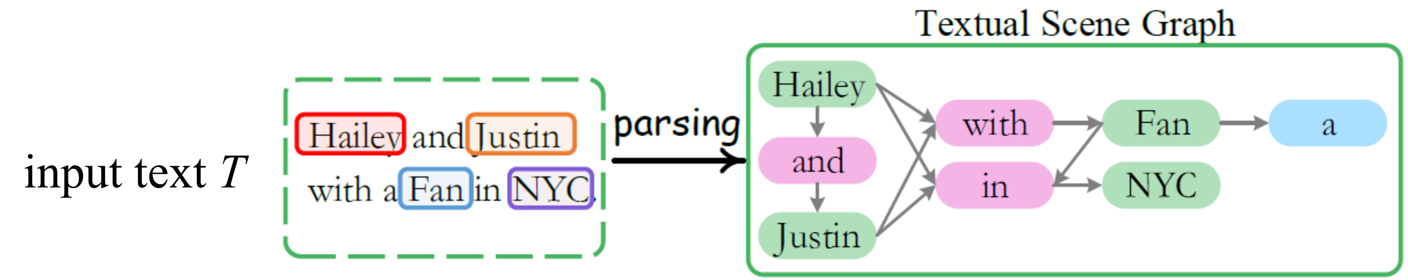
Inference

SG-based Multimodal Relation Extraction

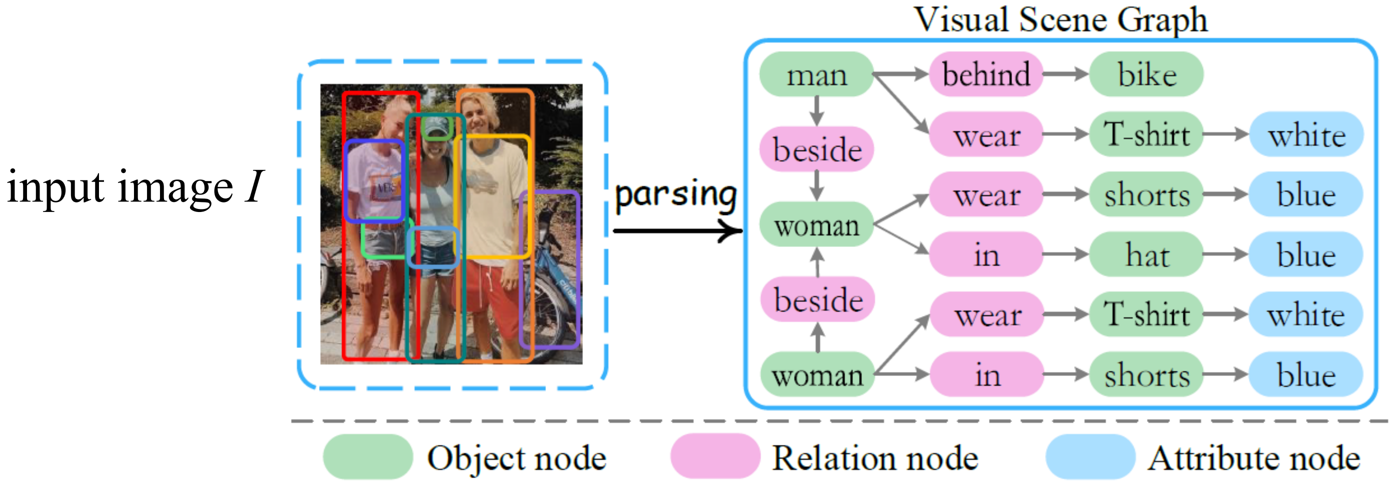
Method

➢ Scene Graph Generation

- Represent input text T with Textual Scene Graph (TSG)



- Represent input image I with Visual Scene Graph (VSG)



SG-based Multimodal Relation Extraction

Method

➢ Cross-modal Graph Construction

- Merge the VSG and TSG into one unified backbone cross-modal graph (CMG)

$$G = (V^T \cup V^I, \boxed{E^T \cup E^I} \cup \boxed{E^{\times}}) \qquad \boldsymbol{X} = \boldsymbol{X}^T \cup \boldsymbol{X}^I$$

intra-modal hyper-edges

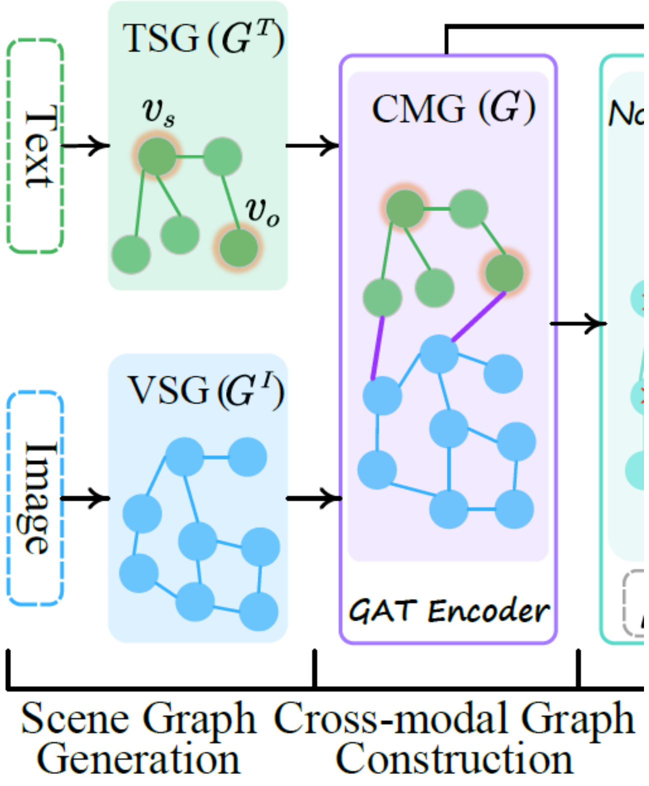
inter-modal hyper-edges

- Creating inter-modal hyper-edges by measuring the relevance score

$$s_{v_i^I, v_j^T} = cos(\boldsymbol{x}_i^I, \boldsymbol{x}_j^T)$$

- Graph Encoding

$$\boldsymbol{H} = \{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_{m+n}\} = \mathrm{GAT}(G, \boldsymbol{X})$$



SG-based Multimodal Relation Extraction

Method

➢ GIB-guided Feature Refinement

- Screen the initial CMG structure i.e., fine-grainedly prune the input image and text features

  - Node Filtering

    Filter out those task-irrelevant nodes

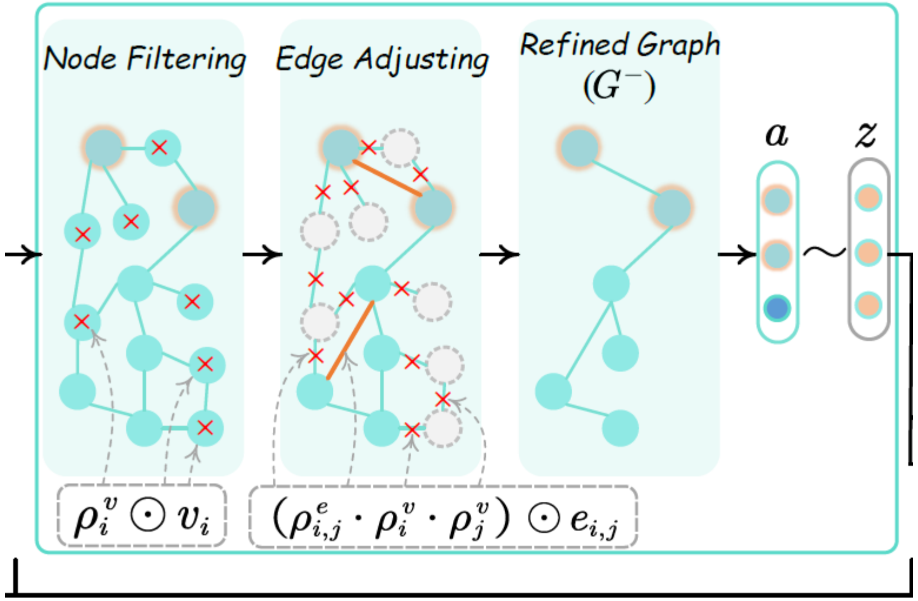
  - Edge Adjusting

    Adjust the edges based on their relatedness to the task inference.

  - GIB-guided optimization

    To ensure that the above adjusted graph G^{-} is sufficiently informative (i.e., not wrongly pruned)

$$\mathcal{L}_{\text{GIB}} = \min_{z} \left[-I(z, Y) + \beta \cdot I(z, G) \right]$$



GIB-guided Feature Refinement

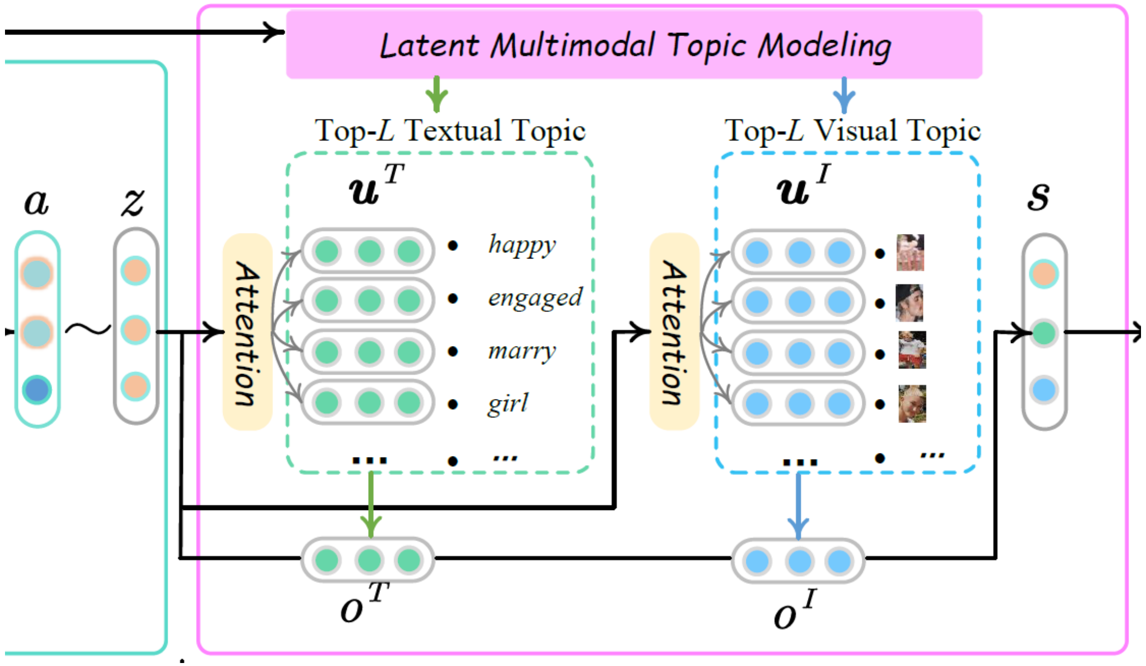
Method

➢ Multimodal Topic Integration

- Enrich the compressed CMG features with more semantic contexts, i.e., the multimodal topic features.
  - Retrieve the associated top-L textual and visual topic keywords
  - Devise an attention operation to integrate the embeddings of the multimodal topic words

$$\alpha_i^{T/I} = \frac{\exp(\text{FFN}([\boldsymbol{u}_i^{T/I}; \boldsymbol{z}]))}{\sum_i^L \exp(\text{FFN}([\boldsymbol{u}_i^{T/I}; \boldsymbol{z}]))},$$

$$\boldsymbol{o}^{T/I} = \sum_i^L \alpha_i^{T/I} \boldsymbol{u}_i^{T/I}.$$



SG-based Multimodal Relation Extraction

Experiment

Main Results

- *Our model achieves the best performance.*
- *Information screening and exploiting both contribute to the task performance.*
- *Scene graph is beneficial for structural modeling of the multimodal inputs.*

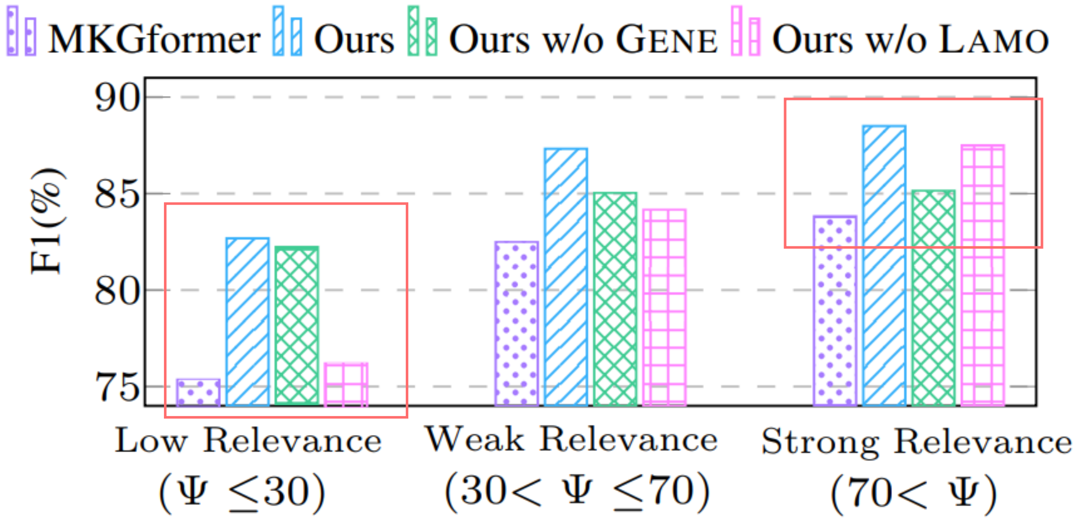
	Acc.	Pre.	Rec.	F1
• *Text-based Methods*				
BERT†	-	63.85	55.79	59.55
PCNN†	72.67	62.85	49.69	55.49
MTB†	72.73	64.46	57.81	60.86
DP-GCN♭	74.60	64.04	58.44	61.11
• *Multimodal Methods*				
BERT(Text+Image)♭	74.59	63.07	59.53	61.25
BERT+SG†	74.09	62.95	62.65	62.80
MEGA†	76.15	64.51	68.44	66.41
$\text{VisualBERT}^{\dagger}_{base}$	-	57.15	59.48	58.30
$\text{ViLBERT}^{\dagger}_{base}$	-	64.50	61.86	63.16
RDS†	-	66.83	65.47	66.14
HVPNeT†	-	83.64	80.78	81.85
MKGformer†	92.31	82.67	81.25	81.95
Ours	**94.06**	**84.69**	**83.38**	**84.03**
w/o GENE (Eq. 11)	92.42	82.41	81.83	82.12
w/o $I(z, G)$ (Eq. 13)	93.64	83.61	82.34	82.97
w/o LAMO (Eq. 4)	92.86	82.97	81.22	82.09
w/o o^T	93.05	83.95	82.53	83.23
w/o o^I	93.63	84.03	83.18	83.60
w/o VSG&TSG	93.12	83.51	82.67	83.09
w/o CMG	93.97	84.38	83.20	83.78

SG-based Multimodal Relation Extraction

Experiment

➢ Analysis and Discussion

Q: *Under what circumstances do the internal-information screening and external-information exploiting help?*



- For the inputs with higher text-vision relevance, the GENE plays a greater role than LAMO, while under the case with less cross-modal feature relevance, LAMO contributes more significantly than GENE.

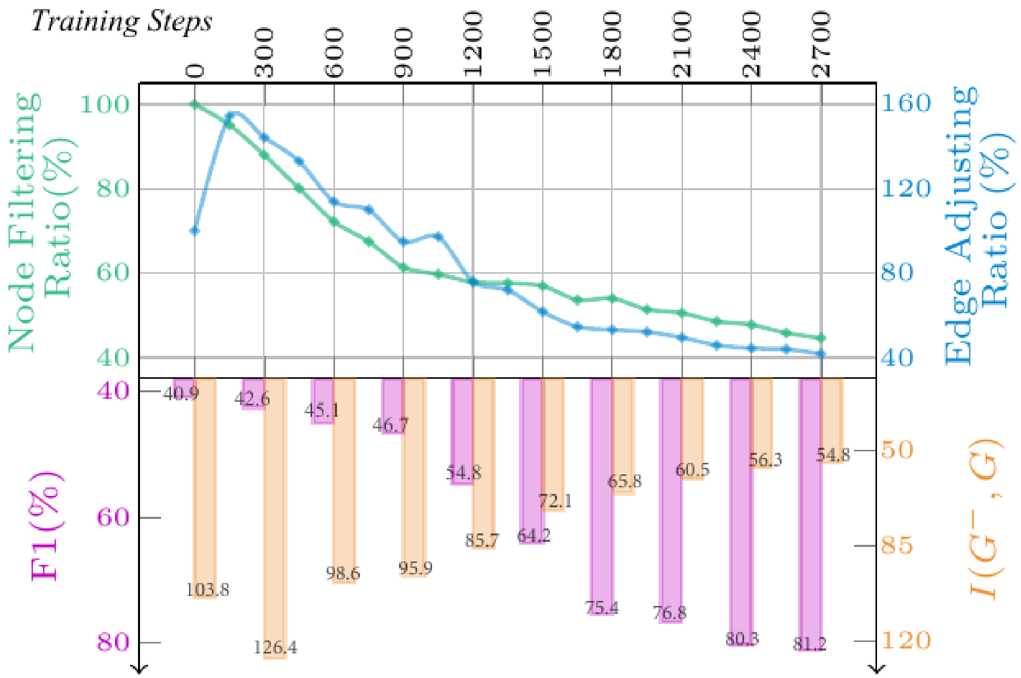
GENE - GIB-guided Feature Refinement
LAMO – Latent Multimodal Topic Model

SG-based Multimodal Relation Extraction

Experiment

➢ Analysis and Discussion

Q: *Does GENE really helps by denoising the input features?*



- Clear pruning pattern.
- Effective performance increase.

SG-bas

Exper

Topic	Textual keywords	Visual keywords (ID)
#Politic	*trump, president, world, new, china, leader, summit, meet, korean, senate*	*#1388, #1068*
#Music	*tour, concert, video, live, billboard, album, styles, singer, taylor, dj*	*#1446, #1891*
#Love	*wife, wedding, engaged, ring, son, baby. girl, love, rose, annie*	*#434, #1091*
#Leisure	*photo, best, beach, lake, island, bridge, view, florida, photograph, great*	*#679, #895*
#Idol	*metgala, hailey, justin, taylor, rihanna, hit, show, annual, pope, shawn*	*#1021, #352*
#Scene	*contain, near, comes, american, in, spotted, travel, to, from, residents*	*#535, #167*
#Sports	*team, man, world, cup, nike, nba, football, join, play, chelsea*	*#1700, #109*
#Social	*google, retweet, twitter, youtube, netflix, acebook, flight, butler, series, art*	*#1043, #1178*
#Show	*show, presents, dress, interview, shot, speech, performing, attend, portray, appear*	*#477, #930*
#Life	*good, life, please, family, dog, female, people, boy, soon, daily*	*#613, #83*

Table 3: Top 10 key textual topic keywords and top 2 visual topic keywords discovered by LAMO.

SG-based Multimodal Machine Translation

Application III:

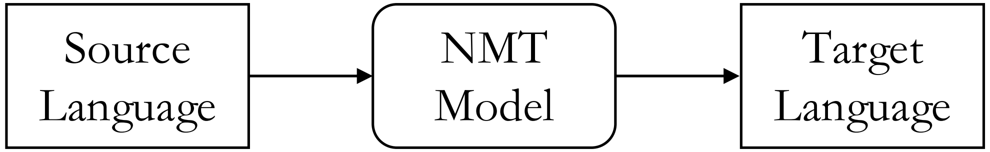
***Scene Graph as Pivoting*: Inference-time Image-free Unsupervised Multimodal Machine Translation with Visual Scene Hallucination**

[1] Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, Tat-Seng Chua. Scene Graph as Pivoting: Inference-time Image-free Unsupervised Multimodal Machine Translation with Visual Scene Hallucination. ACL. 2023.

SG-based Multimodal Machine Translation

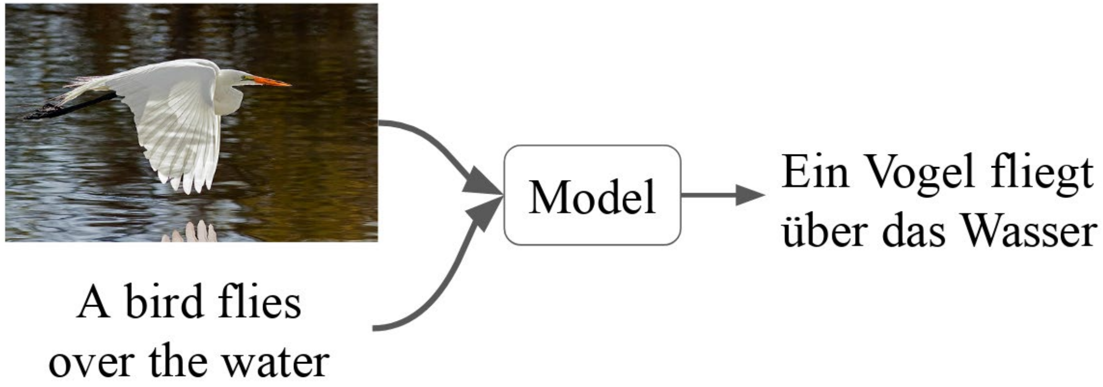
Motivation

➤ Neural Machine Translation (NMT)



- Training: <src-tgt>

➤ Multimodal Machine Translation (MMT)



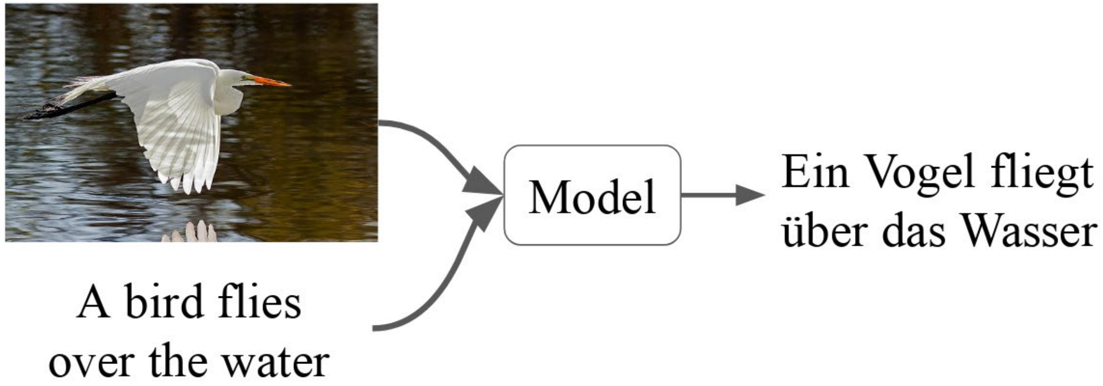
- Training: <src-img-tgt>

Collecting large-scale parallel sentences are cost & sometime infeasible !

SG-based Multimodal Machine Translation

Motivation

- Unsupervised Multimodal Machine Translation (UMMT)



- Training & Testing: <src-img-~~tgt~~> <text-img>

SG-based Multimodal Machine Translation

Motivation

➢ Unsupervised Multimodal Machine Translation (UMMT)

Practical UMMT requires the avoidance of not only parallel sentences during training, but also the paired image during inference (testing).

- ☐ *some existing MMT researches exempt the testing-time visual inputs;*
- ☐ *they all unfortunately are supervised methods, relying on large-scale parallel sentences for training.*

	Avoid parallel sent. during training?	**Avoid paired img. during testing?**
• *Supervised MMT*		
General MMT	×	×
Zhang et al. (2020) Fang and Feng (2022) Li et al. (2022)	×	✓
• *Unsupervised MMT*		
Chen et al. (2018) Su et al. (2019) Huang et al. (2020)	✓	×
This work	✓	✓

Table 1: Practical unsupervised MMT requires the avoidance of not only parallel sentences during training, but also the paired image during inference (testing).

✓ It's necessary to explore the Inference-time Image-free UMMT!

SG-based Multimodal Machine Translation

Motivation

- Unsupervised Multimodal Machine Translation (UMMT)

  - Visual information is vital to UMMT, however both the existing supervised and unsupervised MMT suffer from ineffective and insufficient modeling of visual pivot features.

    - Coarse-grained vision-language alignment learning.

    - Phrase-level vision-language alignment learning (grounding).

✓ Still fail to have a holistic understanding of the visual scene!

SG-based Multimodal Machine Translation

Method

➢ Scene Graph-based UMMT System

- The input src text and paired image are first transformed into LSG and VSG.
- LSG and VSG are further fused into a mixed SG, and then translated into the tgt-side LSG.
- And the tgt sentence will be finally produced conditioned on the tgt LSG.

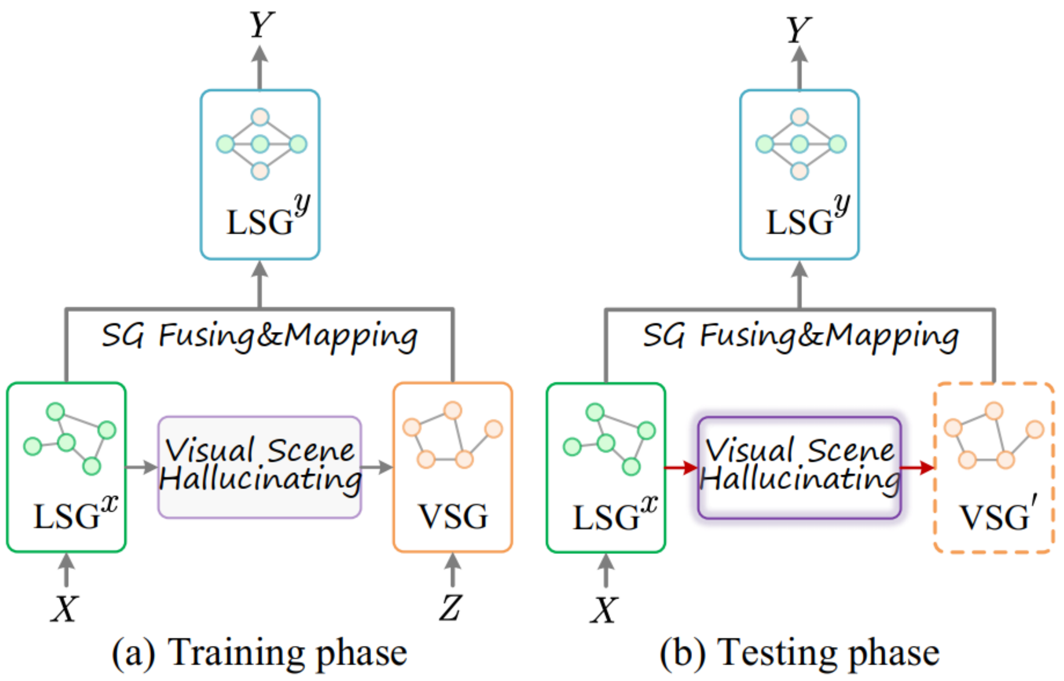


Figure 2: The high-level overview of our SG-based UMMT model. During training, src-side sentences with paired images are used as inputs, together with the corresponding LSG and VSG. Testing phase only takes src-side sentences, where the visual hallucination module is activated to generate VSG from text sources.

SG-based Multimodal Machine Translation

Method

- Visual Scene Hallucination

  - To support pure-text (image-free) input during inference, we devise a novel **visual scene hallucination (VSH)** module.

  - VSH dynamically generates a hallucinated VSG from the LSG compensatively.

    - *Step1: sketching skeleton*

    - *Step2: completing vision*

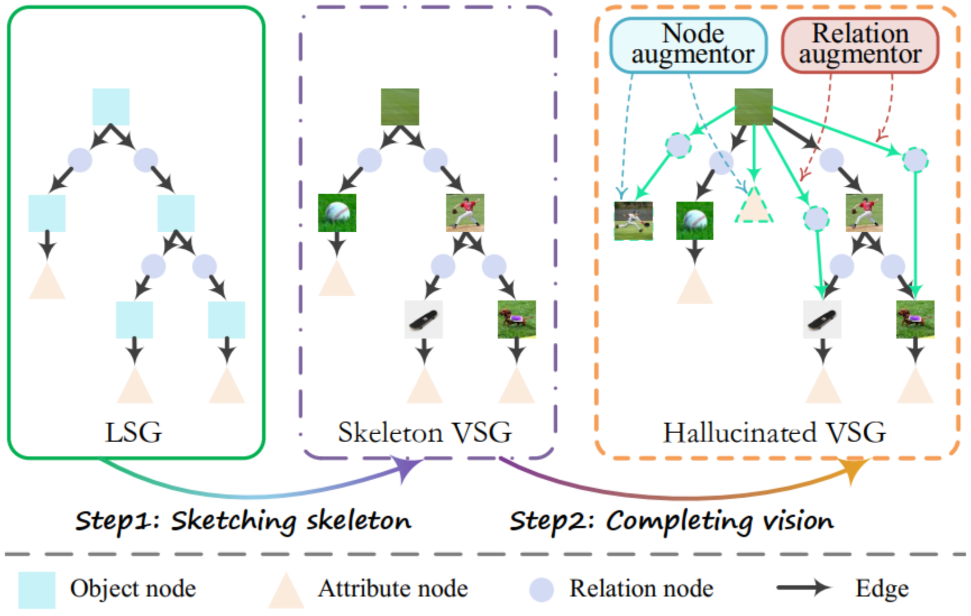


Figure 3: The illustration of the visual scene hallucination (VSH) module, including two steps of inference.

SG-based Multimodal Machine Translation

Method

- Scene Graph Pivoting Learning for UMMT

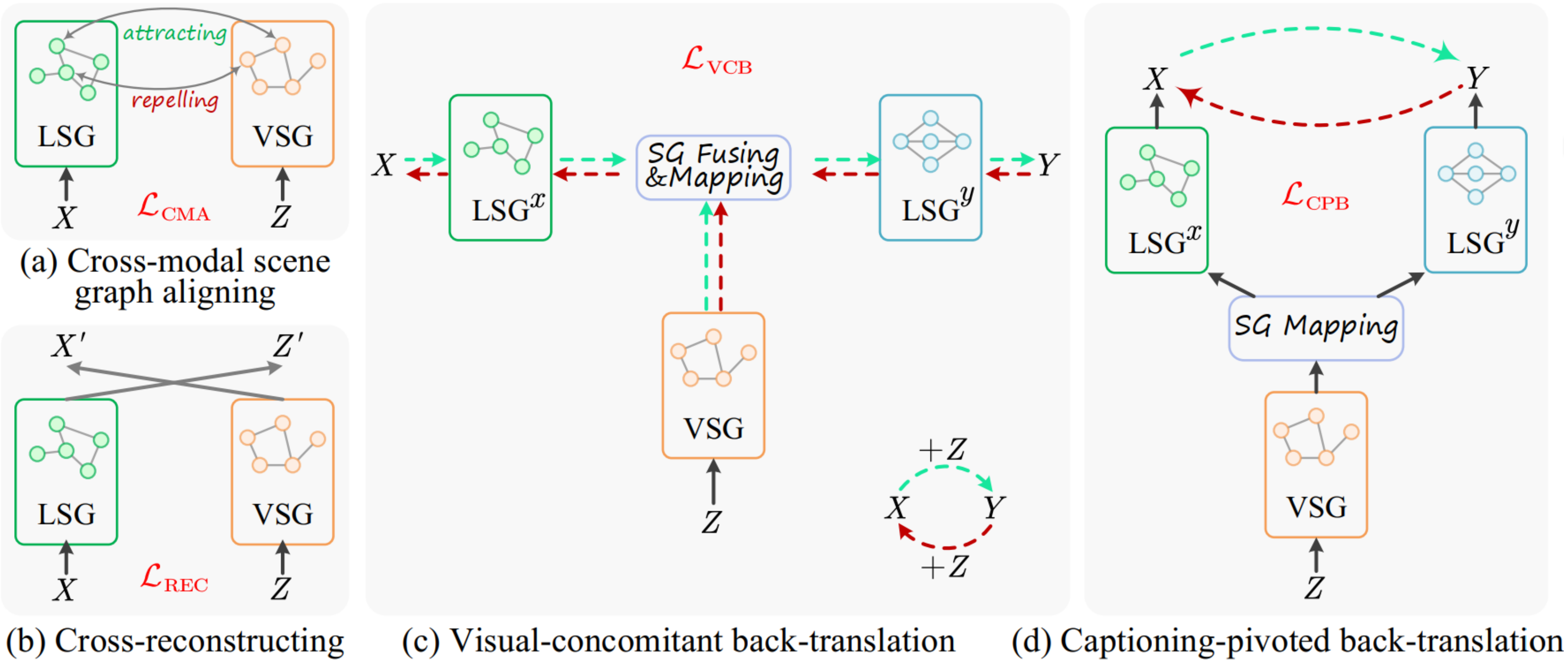


Figure 4: Illustrations of the learning strategies for unsupervised multimodal machine translation.

SG-based Multimodal Machine Translation

Experiment

➢ Main Results

	En → Fr		En ← Fr		En → De		En ← De	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
• ***Testing with image input given***								
Game-MMT	-	-	-	-	16.6	-	19.6	-
UMMT	39.8	35.5	40.5	37.2	23.5	26.1	26.4	29.7
PVP	52.3	67.6	46.0	39.8	33.9	54.1	36.1	34.7
Ours$^{\#}$	**56.9**	**70.7**	**50.4**	**42.5**	**37.4**	**57.2**	**39.2**	**38.3**
w/o SGs	51.7	64.0	46.2	40.7	34.5	56.4	36.9	35.2
• ***Testing without image input given***								
UMMT	15.8	12.7	10.2	13.6	8.4	11.3	7.5	10.8
UMMT*	30.4	28.4	31.8	30.4	15.7	17.7	19.3	22.7
PVP	26.1	23.8	25.7	23.4	11.1	13.8	14.0	17.2
PVP*	46.7	58.0	39.0	31.9	25.4	40.1	27.6	26.0
Ours	**50.6** (+3.9)	**64.7** (+6.7)	**45.5** (+6.5)	**37.3** (+5.4)	**32.0** (+6.6)	**52.3** (+12.2)	**33.6** (+6.0)	**32.8** (+6.8)

Table 2: Results of UMMT on Multi30K data. 'Ours$^{\#}$': using paired images for testing instead of visual hallucination. 'UMMT*/PVP*': re-implemented baselines with phrase-level retrieval-based visual hallucination. In the brackets are the improvements of our model over the best-performing baseline(s).

Our system shows significant improvements over the best baseline PVP*, by average **5.75**=(3.9+6.5+6.6+6.0)/4 BLEU score.

SG-based Multimodal Machine Translation

Experiment

- **The longer and more complex the sentences, the higher the translation quality benefiting from the SGs features.**

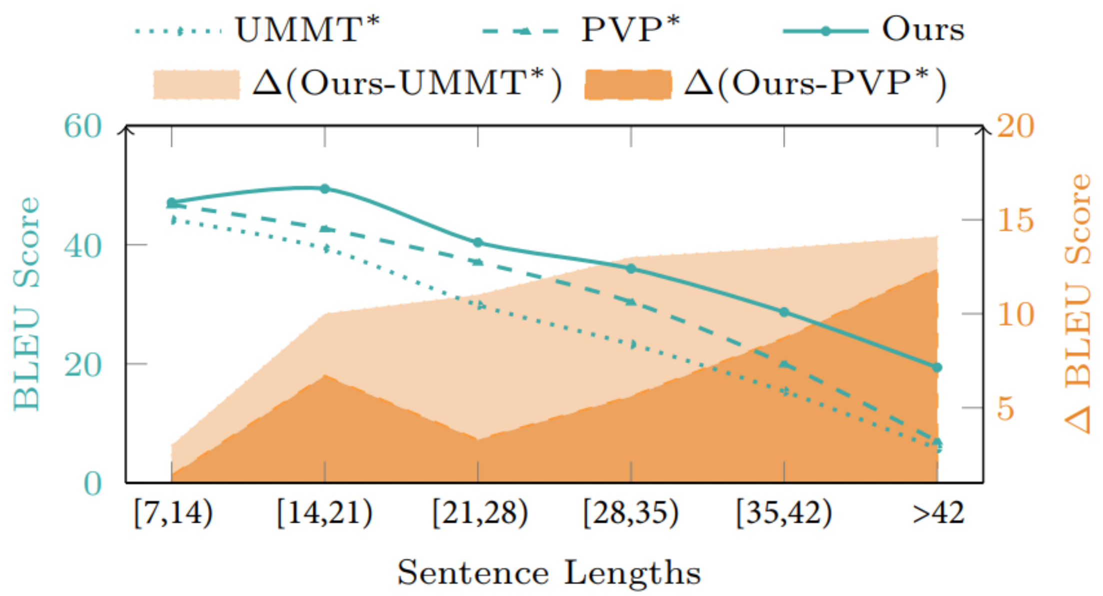


Figure 6: BLEU scores under different sentence lengths.

SG-based Multimodal Machine Translation

Experiment

➢ **SG-based visual scene hallucination mechanism helps gain rich and correct visual features.**

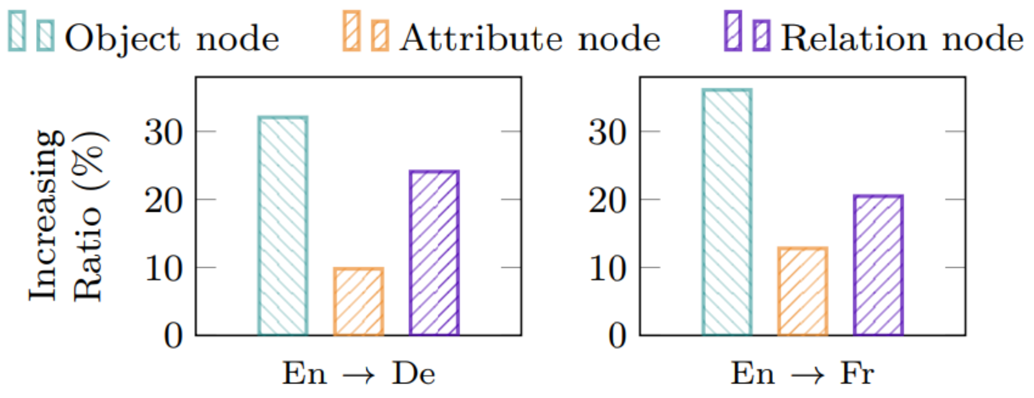


Figure 7: Growing rate of nodes in hallucinated VSG.

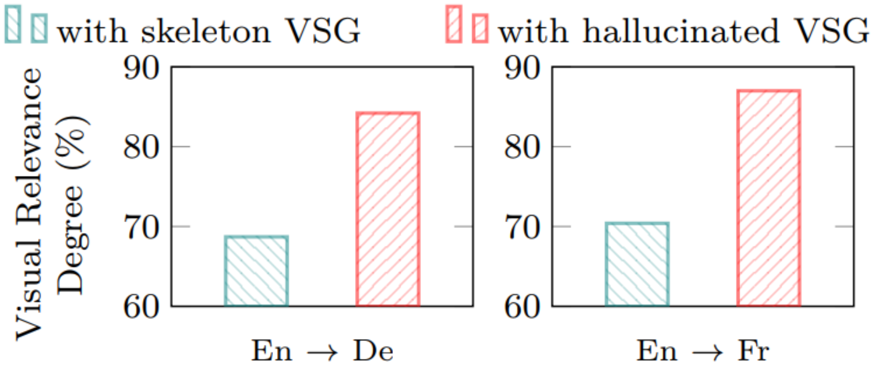


Figure 8: Degree of visual relevance (similarity) between the hallucinated vision (via graph-to-image generator) and the ground truth image.

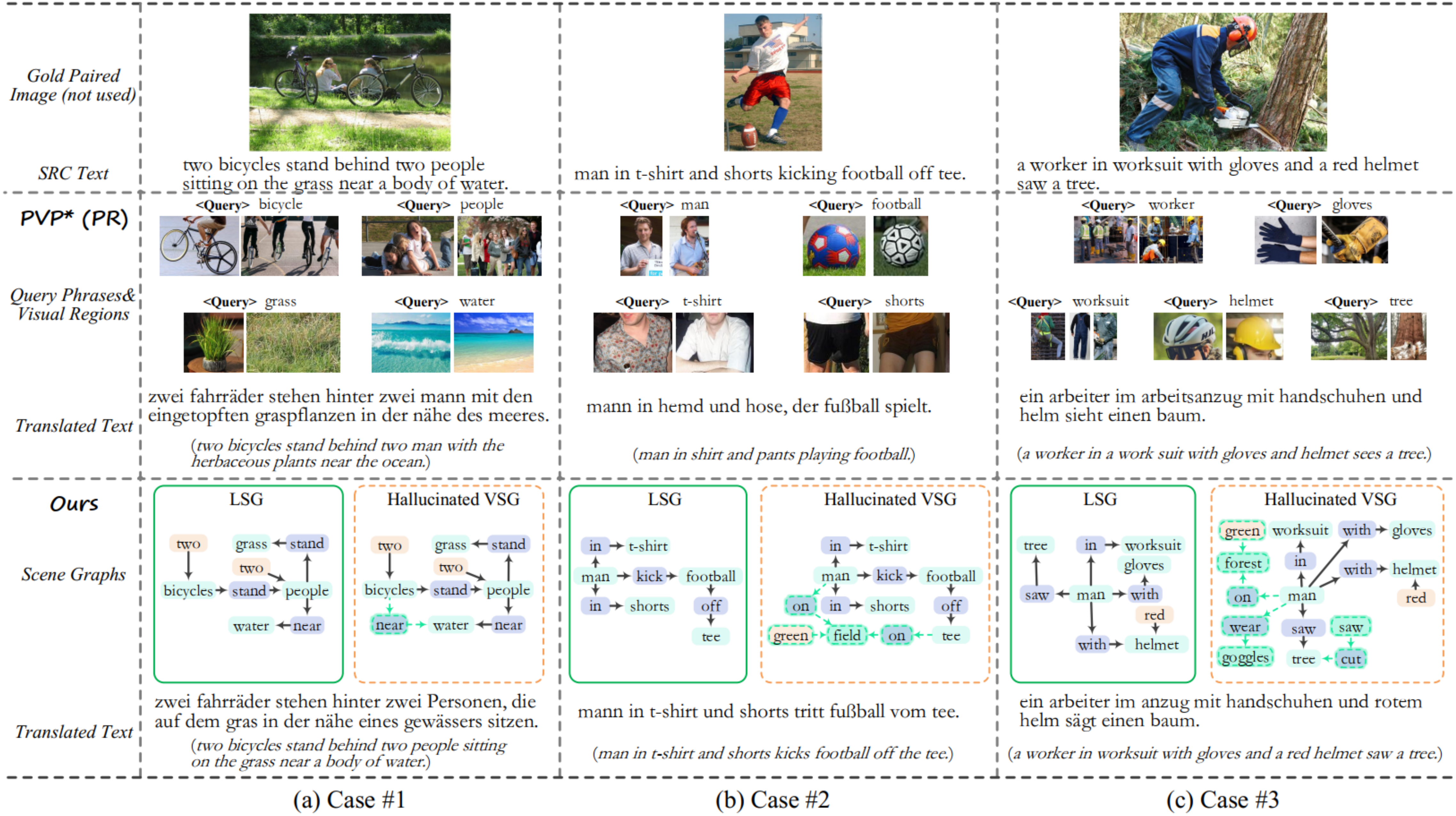


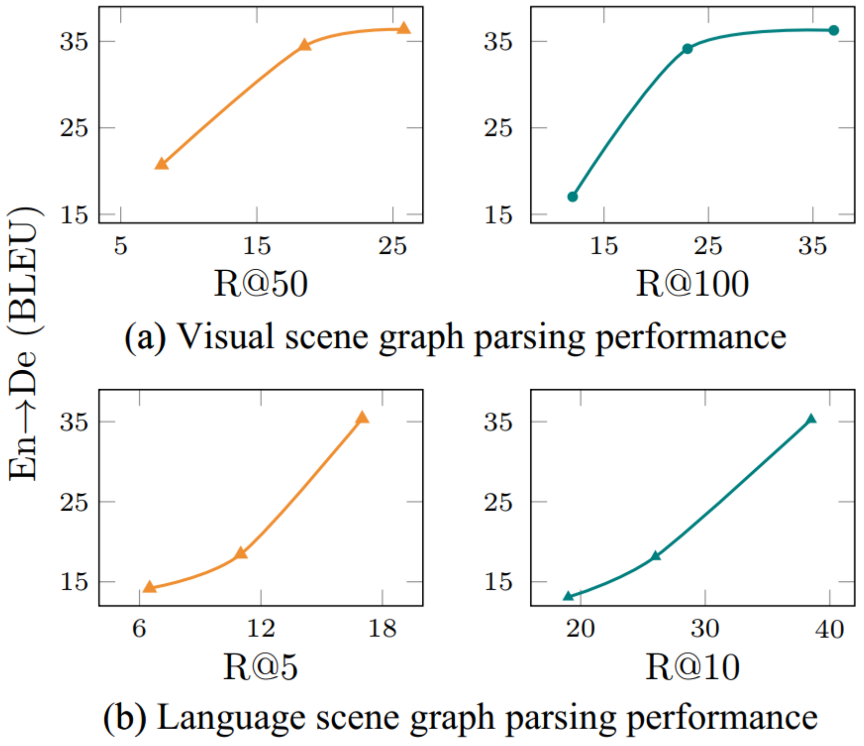
Figure 5: Qualitative results of inference-time image-free UMMT (En→De).

SG-based Multimodal Machine Translation

Experiment

➤ To what extent does SG parsing quality influence the efficacy of end task?

- Surely low-quality SG annotations decrease the efficacy of the SG features for end tasks.
- Existing SoTA SG parsers are effective enough to aid the end-tasks, i.e., the positive outweighs negative.
- Mostly, end-tasks are more sensitive to the quality of the textual SG, compared with the visual SG.



CONTENT

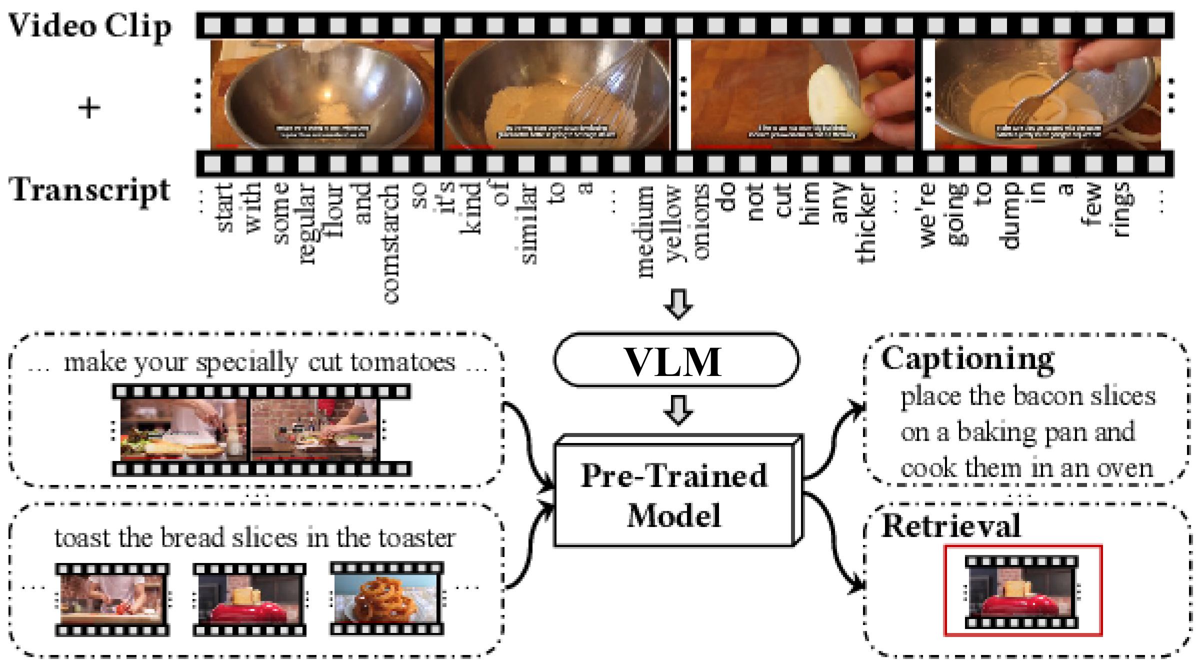
Application IV:

Enhancing Video-Language Representations with Structural Spatio-Temporal Alignment

[1] Hao Fei, Shengqiong Wu, Meishan Zhang, Shuicheng YAN, Min Zhang, Tat-Seng Chua. Enhancing Video-Language Representations with Structural Spatio-Temporal Alignment. 2023.

Motivation

- Video-language model (VLM) pre-training
  - ✓ Existing issues:
    - *Coarse-grained cross-model aligning*
    - *Under-modeling of temporal dynamics*
    - *Detached video-language view*

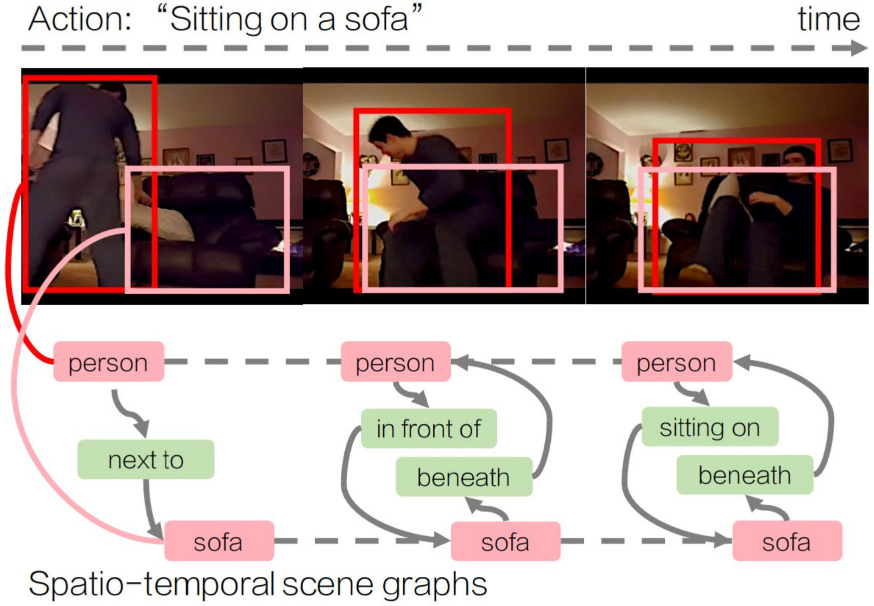


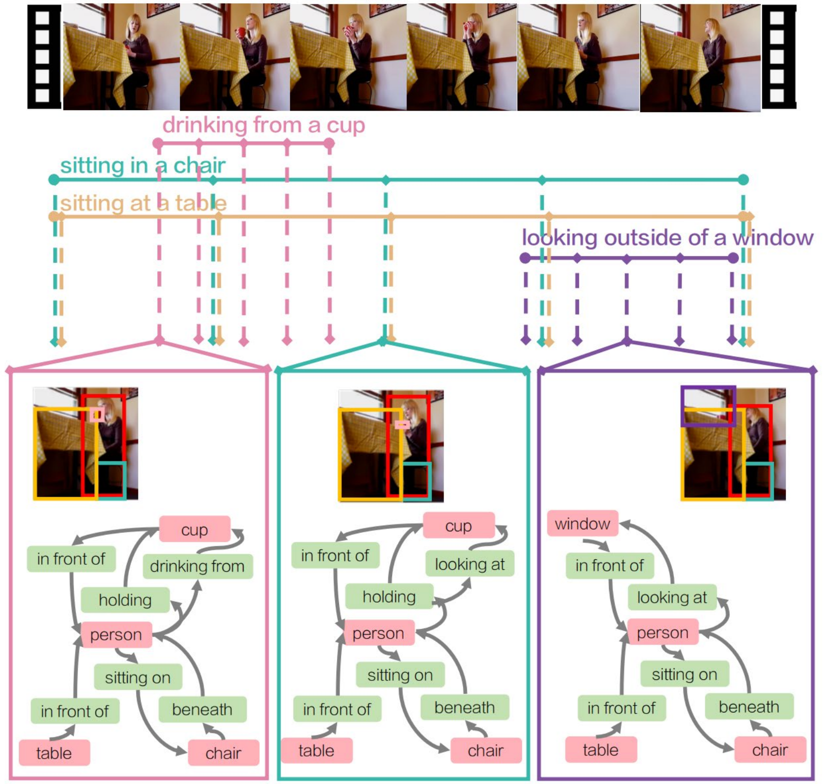
SG-based Video-Language Modeling

Video Scene Graph Representation

- Video Scene Graph, aka., Dynamic Scene Graph (DSG), Spatio-temporal Scene Graph

A sequence of VSG along time frames.





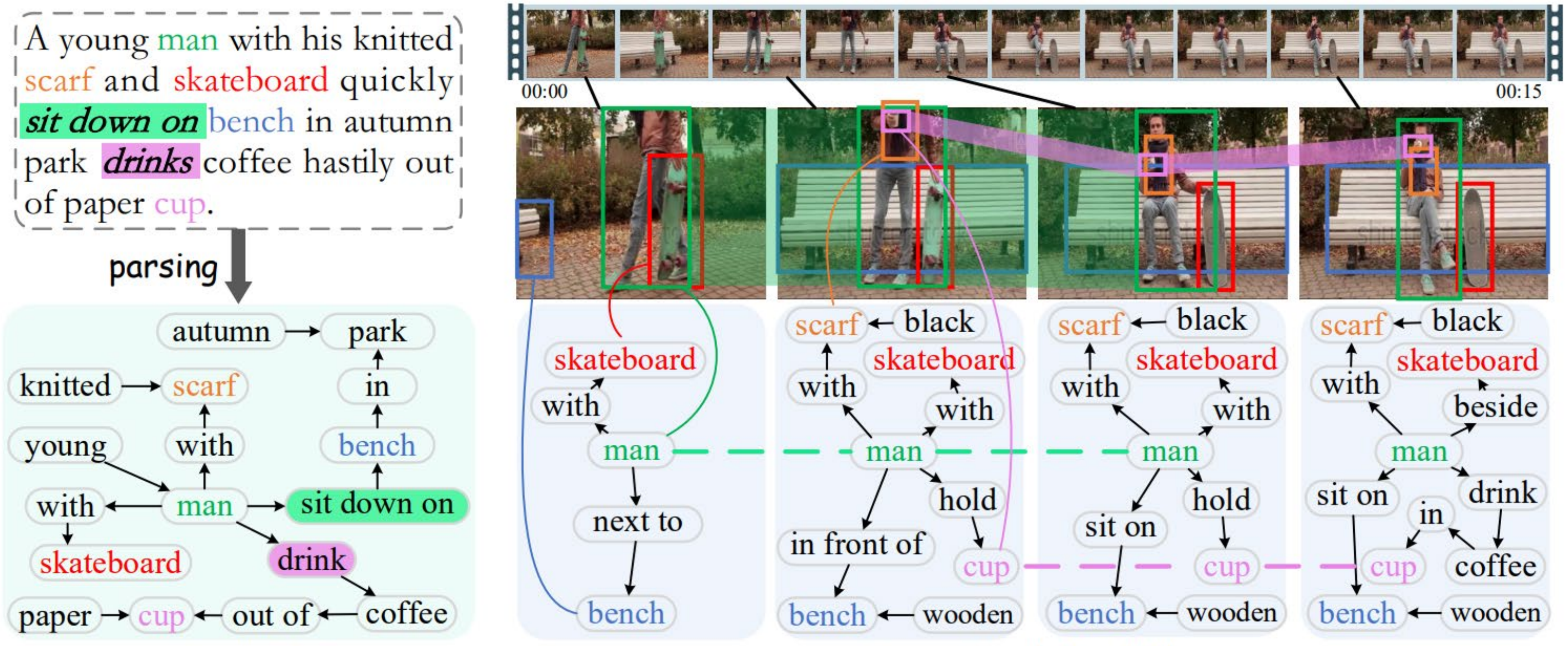
[1] *Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. CVPR, 2020.*

SG-based Video-Language Modeling

Motivation

- video-language model (VLM) pre-training

- *Coarse-grained cross-model aligning*
    - *Fine-grained alignment*
- *Under-modeling of temporal dynamics*
    - *Modeling dynamics with DSG*
- *Detached video-language view*
    - *Merging TSG and DSG*



Scene Graph for Text

Scene Graph for Video

Method

➢ **Fine-grained Structural Spatio-temporal Alignment (Finsta)** framework

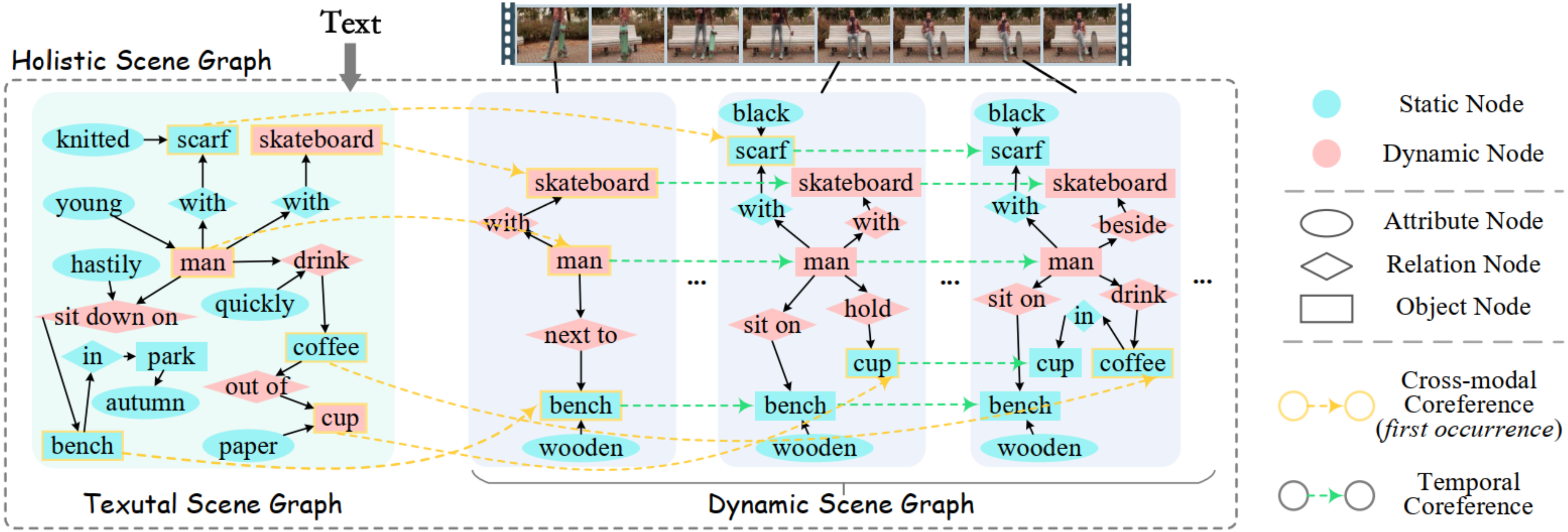


Figure 2: We represent the input text and video with textual scene graph (TSG) and dynamic scene graph (DSG), respectively, where all nodes are categorized into the static type and dynamic type. We further unify the TSG and DSG into a holistic SG (HSG).

SG-based Video-Language Modeling

Method

- **Finsta**
  - SG Representation Construction
    - TSG
    - DSG
    - Holistic SG (HSG)
  - VL Representation Learning
    - Fine-grained Structural Spatio-Temporal Alignment Learning
      - *Object-centered Spatial Contrasting (OSC)*
      - *Predicate-centered Temporal Contrasting (PTC)*
    - Representation Transfer Learning

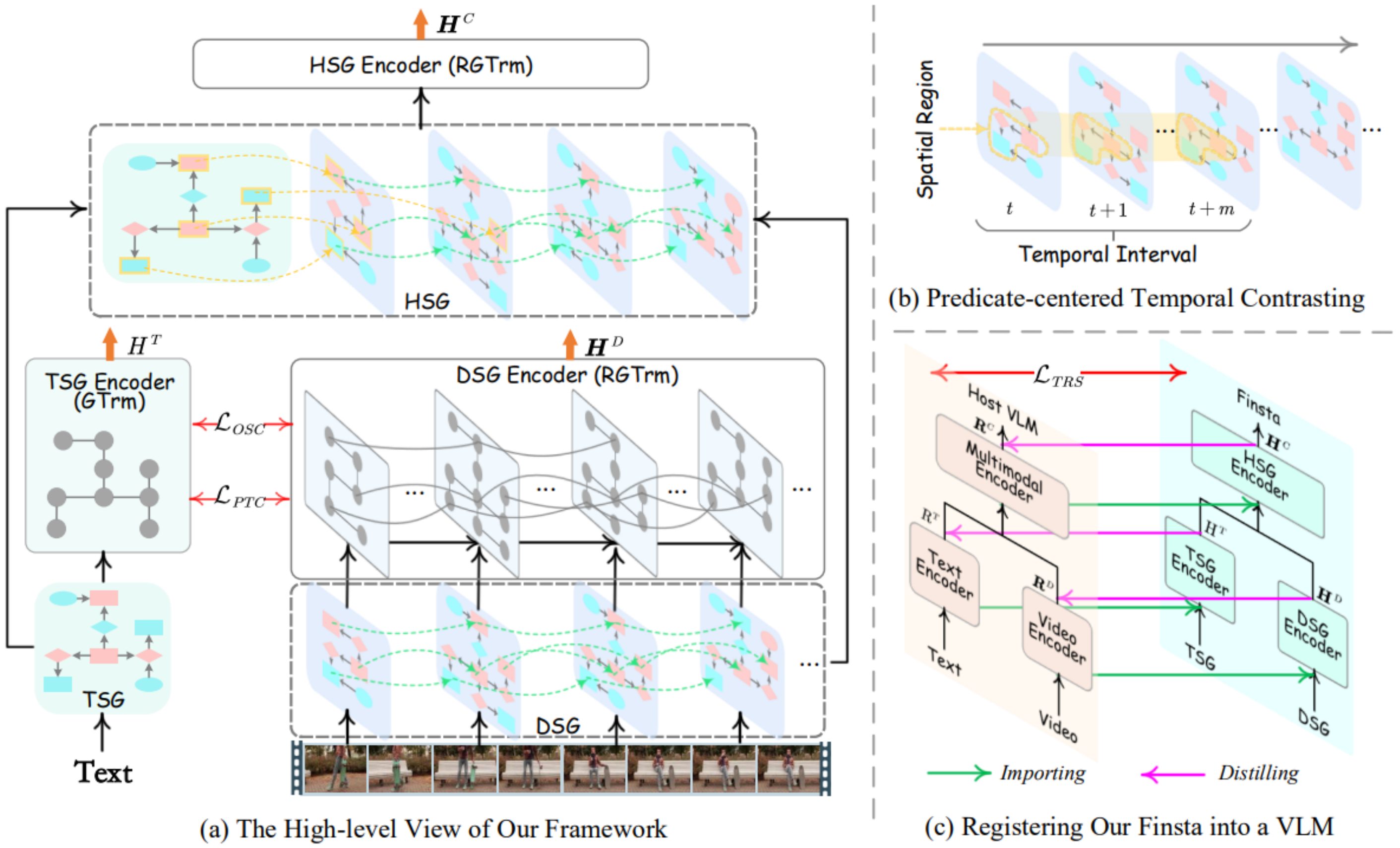


Figure 3: (a) Fine-grained structural spatio-temporal alignment learning (Finsta) based on the dual-stream framework with three SG encoders. (b) Extracting the spatial region and temporal interval for the predicate-centered temporal alignment. (c) Injecting our Finsta representations into a host LVM.

SG-based Video-Language Modeling

Experiment

Table 1: Video Action Recognition results (Acc. on Top-1) on two datasets. The best results are in bold.

Method	K400 [35]	SSV2 [23]
TimeSformer [5]	78.0	59.5
Frozen [3]	78.5	61.6
OmniVL [68]	79.1	62.5
HDVILA	78.6	61.3
Finsta-HDVILA	80.4	63.2
Clover	78.8	62.3
Finsta-Clover	**81.2**	**64.1**

Table 2: Video captioning results on three datasets.

Method	YouCook2 [80]		MSRVTT [72]		MSVD [7]	
	M	B@4	M	B@4	M	B@4
VideoBERT [62]	11.0	4.1	-	-	-	-
UniVL [46]	17.6	11.2	-	-	-	-
SAM-SS [9]	-	-	29.3	45.8	39.0	62.4
SemSynAn [55]	-	-	30.4	46.4	41.9	64.4
OmniVL [68]	14.8	8.7	-	-	-	-
HDVILA	13.5	8.2	32.4	46.0	42.5	64.8
Finsta-HDVILA	**18.8**	**12.7**	36.9	48.6	44.8	66.5
Clover	14.2	9.0	34.1	47.5	43.3	64.6
Finsta-Clover	18.6	12.5	**38.8**	**49.3**	**45.2**	**67.4**

Table 5: Long-Form Video Question-Answering results (Acc.) on two datasets.

Method	How2QA [41]	VIOLIN [44]
ResNet-SF [42]	74.3	-
GVE [10]	-	68.4
HERO [41]	74.3	68.6
LFVILA	76.1	70.9
Finsta-LFVILA (S-Vid)	77.5	71.7
Finsta-LFVILA (L-Vid)	**78.8**	**73.0**

Table 3: Video Question Answering results (Acc.) on two datasets.

Method	MSRVTT [70]	MSVD [70]
ClipBERT [38]	37.4	-
VIOLET [18]	43.9	47.9
ALPRO [39]	42.1	45.9
OmniVL [68]	44.1	51.0
HDVILA	40.0	50.7
Finsta-HDVILA	43.4	53.3
Clover	42.5	51.1
Finsta-Clover	**45.8**	**54.6**

Table 4: Video-Text Retrieval results on two datasets.

Method	LSMDC [48]			DiDeMo [26]		
	R@1	R@5	R@10	R@1	R@5	R@10
OA-Trans [67]	18.2	34.3	43.7	34.8	64.4	75.1
ALPRO [39]	-	-	-	35.9	67.5	78.8
CLIP4CLIP [47]	21.6	41.8	49.8	43.4	70.2	80.6
CAMOE [12]	22.5	42.6	50.9	43.8	71.4	-
HDVILA	21.8	42.3	49.7	45.7	72.4	79.2
Finsta-HDVILA	25.3	46.3	55.8	49.3	75.9	83.6
Clover	24.8	44.0	54.5	50.1	76.7	85.6
Finsta-Clover	**26.9**	**46.8**	**56.3**	**51.0**	**77.8**	**86.4**

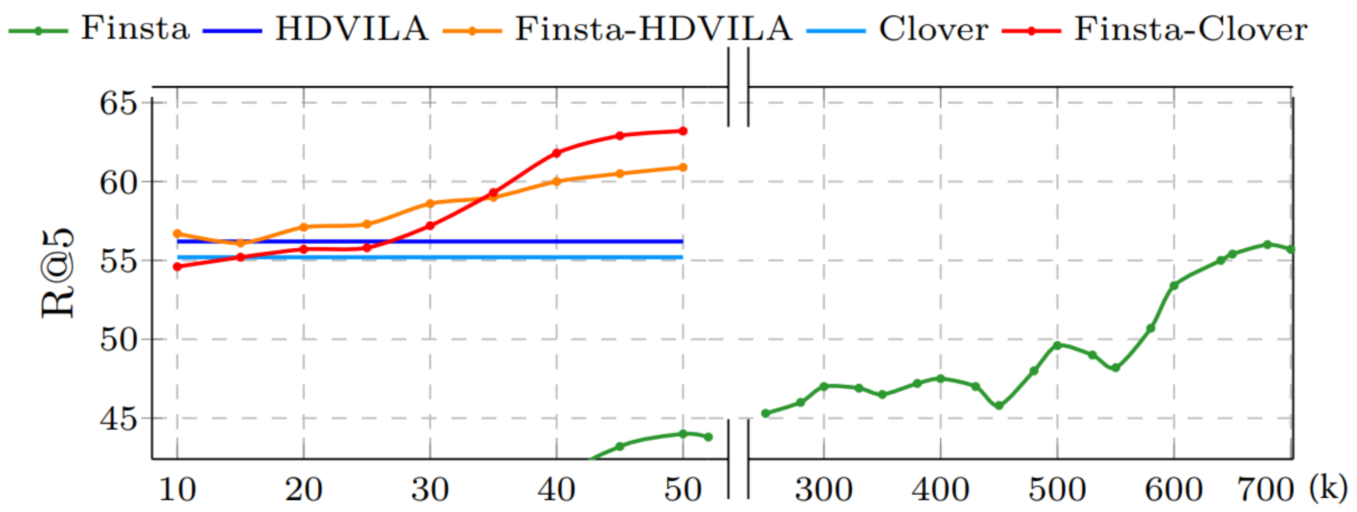
Table 6: Video-Paragraph Retrieval results on QuerYD data [51].

Method	R@1	R@5	R@10
TeachText [14]	14.4	37.7	50.9
Frozen [3]	53.8	75.7	82.7
LFVILA	69.7	85.7	90.3
Finsta-LFVILA (S-Vid)	70.0	86.4	91.2
Finsta-LFVILA (L-Vid)	**73.4**	**87.8**	**93.0**

SG-based Video-Language Modeling

Experiment

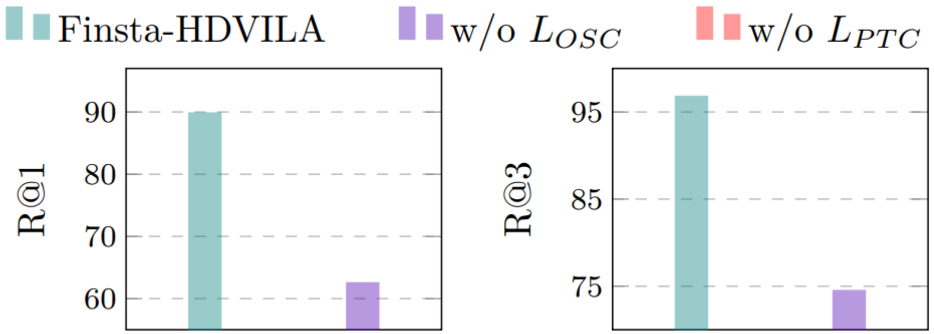
- Influence of Post-training Data Amount



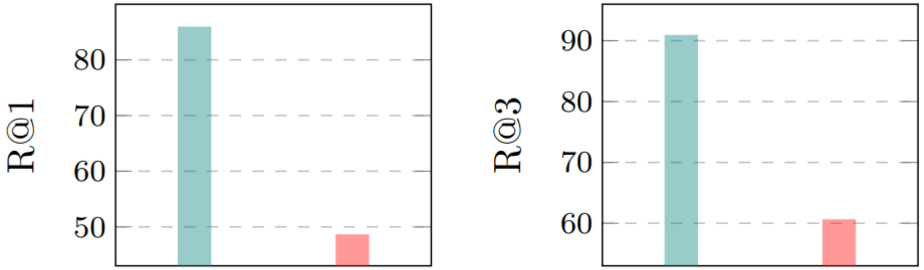
SG-based Video-Language Modeling

Experiment

- Probing Fine-grained Video-Language Correspondences
  - Static Entity-Object Correspondence.
  - Dynamic Predicate-Action Tracking Correspondence.



(a) Static entity-object correspondence



(b) Dynamic predicate-action tracking correspondence

Application V:

Constructing Holistic Spatio-Temporal Scene Graph for Video Semantic Role Labeling

[1] Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, Tat-Seng Chua. Constructing Holistic Spatio-Temporal Scene Graph for Video Semantic Role Labeling. ACM MM. 2023.

SG-based Video Semantic Role Labeling

Motivation

- Video Semantic Role Labeling (VidSRL) *"who does what to whom, where and when and how" within a video*

  - Subtask-1:

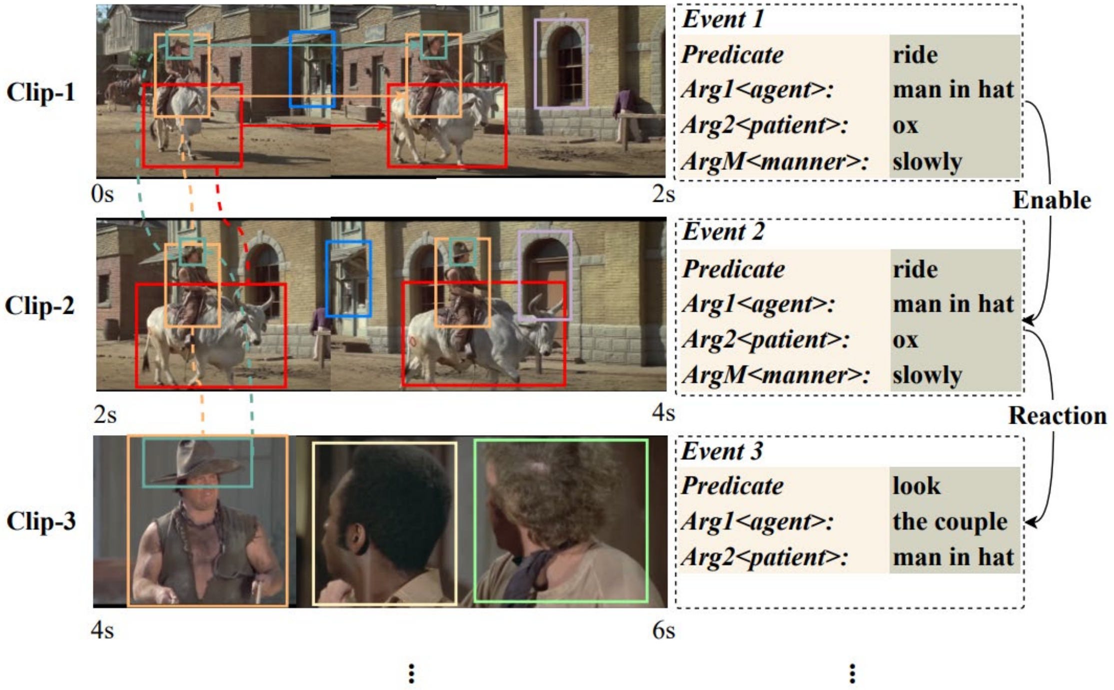
    verb prediction

  - Subtask-2:

    arguments generation (or role labeling)

  - Subtask-3:

    event relation prediction



SG-based Video Semantic Role Labeling

Motivation

➢ Two key bottlenecks in VidSRL

- Lack of fine-grained spatial scene perception
- Insufficient modeling of video temporality

SG-based Video Semantic Rol

Method

➢ Constructing a <u>holistic spatio-temporal scene graph</u> (HostSG)

- Step-1:

  Video dynamic SG (DSG) Generation for Clip.

- Step-2:

  Merging DSG.

- Step-3:

  HostSG Construction.

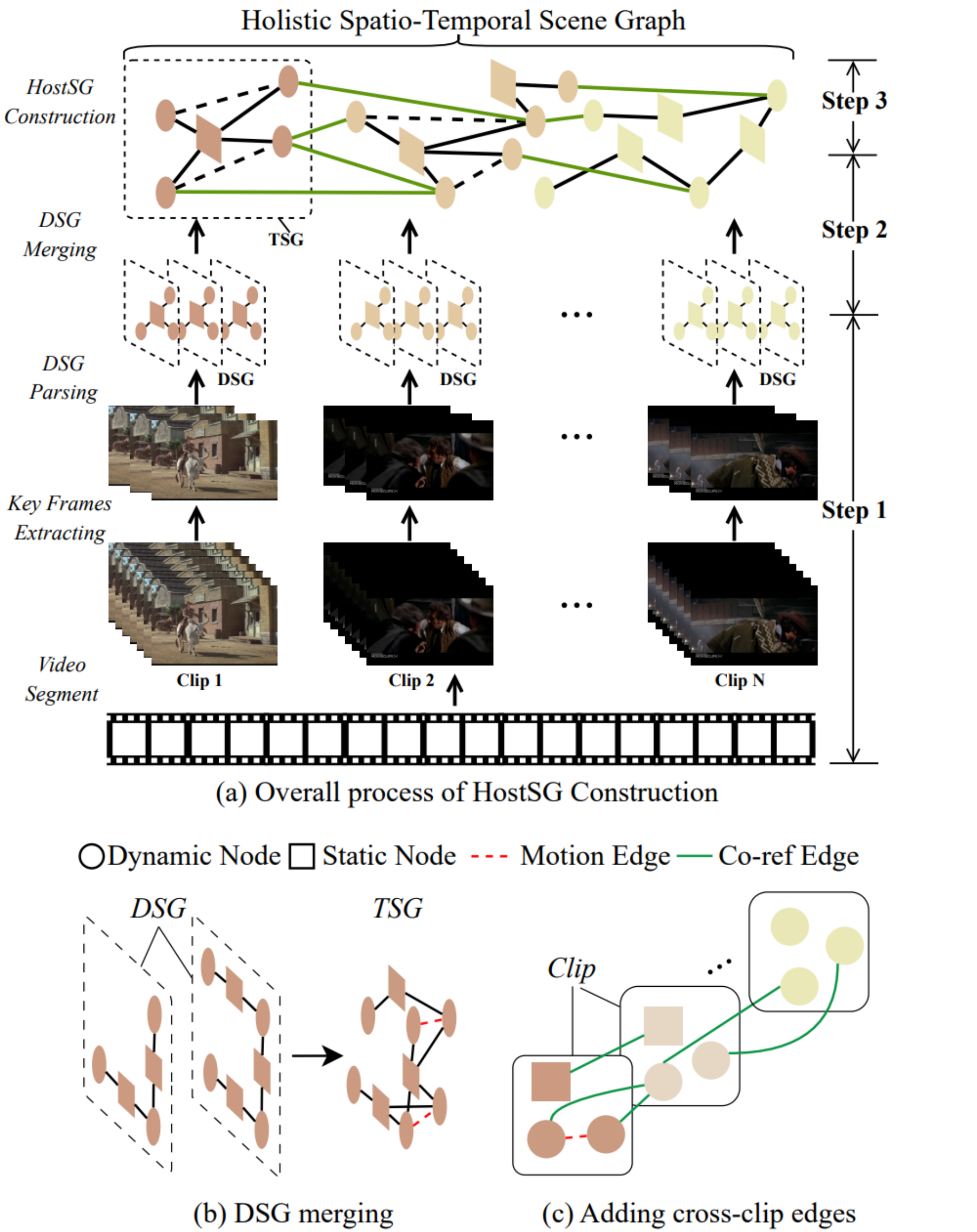


Figure 2: Holistic spatio-temporal scene graph generation.

SG-based Video Semantic Role Labeling

Method

- VidSRL Framework

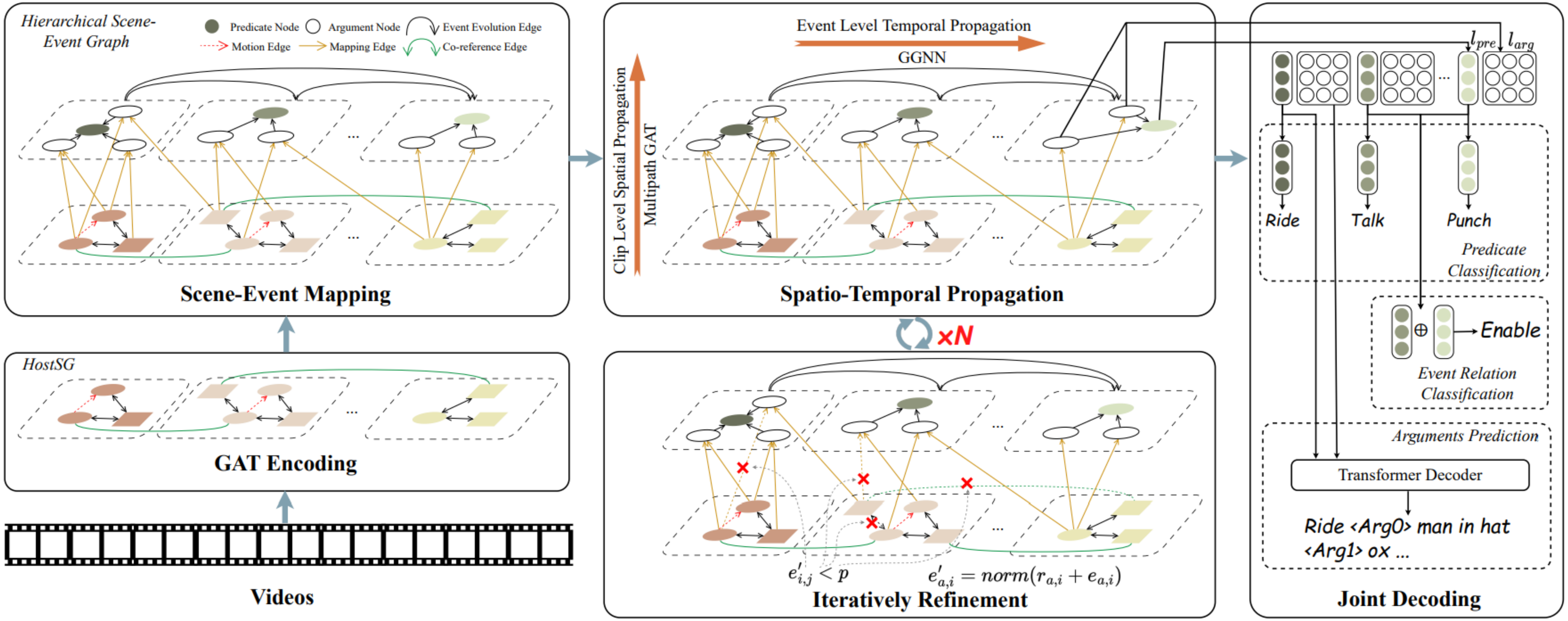


Figure 3: Augmented holistic event-arguments semantic graph.

Experiment

➢ Main Results

Table 1: Main results on the VidSRL dataset. "Verb Cls", "SRL" and "EvtRel" represents the three subtasks verb classification, semantic role labeling and event relation prediction. The CIDEr score is also computed over every verb-sense (CIDEr-Verb) and over argument-types (CIDEr-Arg). Bold numbers are the best, and underlined ones are the second best. Our results are averaged on five running with different seeds. Gray color: methods use ground-truth verb annotations for SRL training.

	Verb Cls			SRL						EvRel
	Acc@1(%)	Acc@5(%)	Rec@5(%)	CIDEr	Rouge-L	CIDEr-Vb	CIDEr-Arg	Lea	Lea-S	Macro-Acc(%)
• Pipeline										
VidSitu-GPT2 [32]	-	-	-	34.67	40.08	42.97	34.45	48.08	28.10	-
VidSitu-I3D [32]	30.17	66.83	4.88	47.06	42.41	51.67	42.76	48.92	33.58	-
VidSitu-SlowFast [32]	32.64	69.20	6.11	45.52	42.66	55.47	42.82	<u>50.48</u>	31.99	<u>34.13</u>
• Joint										
VidSitu-e2e [47]	46.79	75.90	23.38	30.33	29.98	39.56	23.97	35.92	-	-
OME [47]	52.75	83.88	28.44	47.82	40.91	54.51	44.32	-	-	-
OME(disp) [47]	53.32	<u>84.00</u>	28.61	48.46	<u>41.89</u>	56.04	<u>44.60</u>	-	-	-
OME(disp)+OIE [47]	<u>53.36</u>	83.94	<u>28.72</u>	47.16	40.86	53.96	42.78	-	-	-
VideoWhisperer [24]	45.06	75.59	25.25	<u>52.30</u>	35.84	<u>61.77</u>	38.18	38.00	-	-
HostSG (Ours)	**56.15** (+2.79)	**86.33** (+2.33)	**29.38** (+0.66)	**55.09** (+2.79)	**43.13** (+1.24)	**64.24** (+2.47)	**47.68** (+3.08)	**55.70** (+5.22)	**35.01** (+3.2)	**35.97** (+1.84)

Experiment

➤ Q: Does HostSG provide informative spatial and temporal features for VidSRL?

Table 3: Influence of different numbers of frame extraction. 'w/o Key Frame Extraction' means we extract frames with a constant interval.

	Acc@1	CIDEr	Macro-Acc
• 1 Frame/Clip	41.48	36.85	33.91
w/o Key Frame Extraction	41.51	37.10	34.02
• 5 Frames/Clip	**56.15**	**55.09**	**35.97**
w/o Key Frame Extraction	56.13	54.77	35.16
• 11 Frames/Clip	55.15	54.72	35.31
w/o Key Frame Extraction	55.04	54.67	35.29

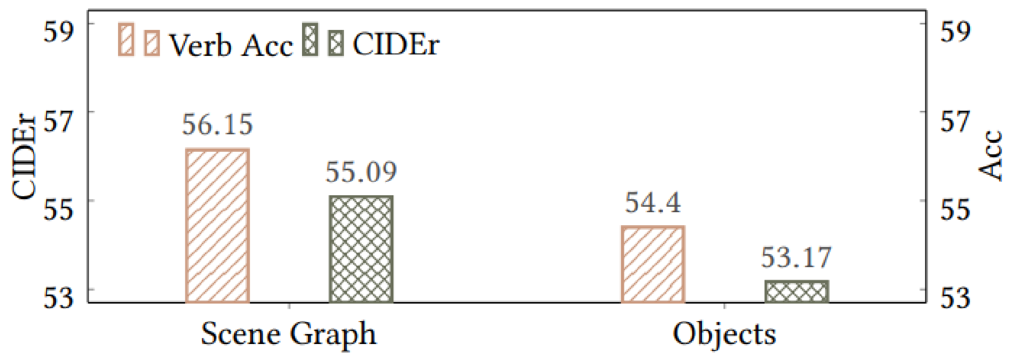


Figure 4: Comparison between the results of scene graph features and object features.

Experiment

➢ Visualization of the cross-clip coreference edges

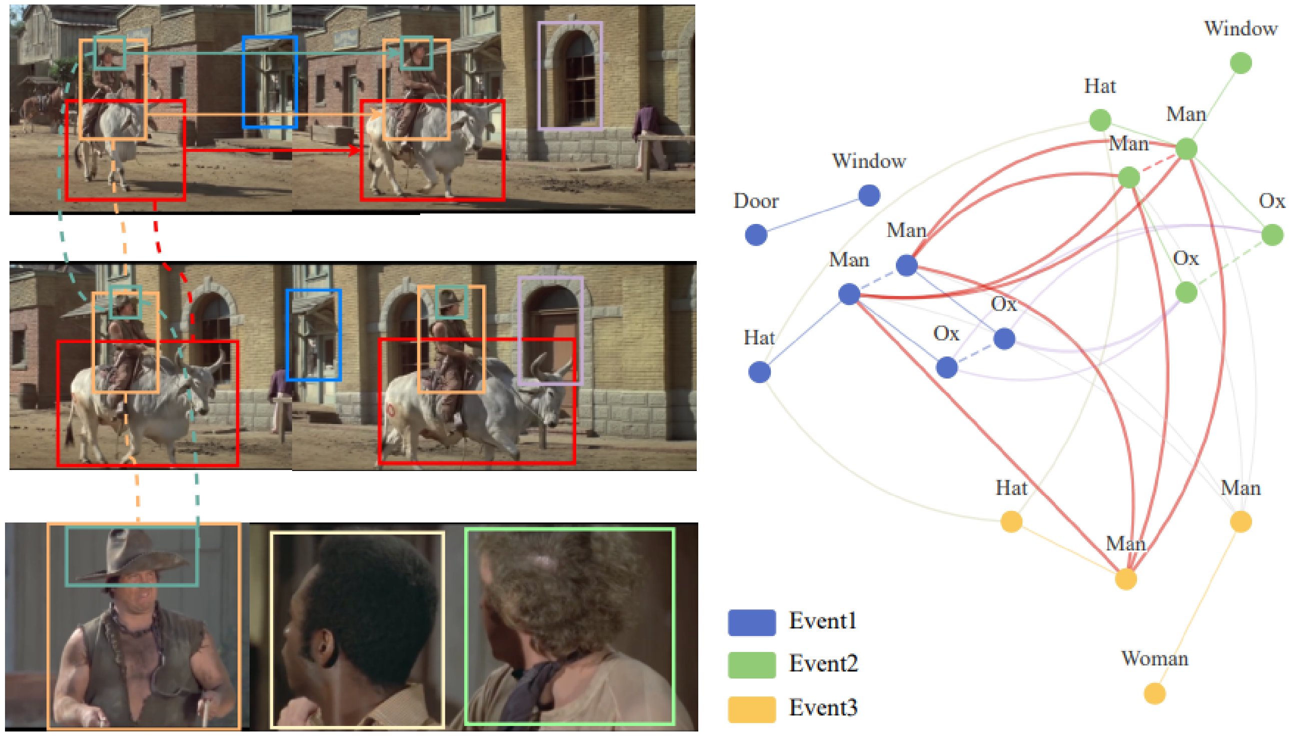
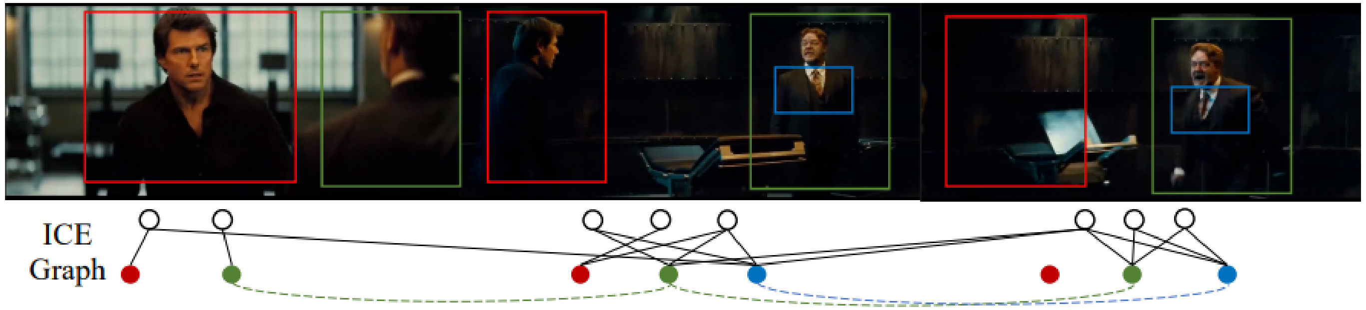


Figure 6: Visualization of the cross-clip coreference edges. We select three clips from a video, representing the edge weights by the line width. The highlighted red lines denote the coreference relation of the objects with the tag "Man".

SG-based Video Semantic Role Labeling

Experiment

➢ Quantitative results



➢ **With SG-Event Mapping**

Event 1: look	*Even 2:* speak	*Event 3:* move
<Arg0> man without tie	<Arg0> man with tie	<Arg0> man with tie
<Arg1> man with tie	<Arg1> man without tie	<ArgM> in anger
<ArgM> across the room	<ArgM> angrily	<ArgM> backwards

➢ **Only HostSG**

Event 1: look	*Event 2:* speak	*Event 3:* look
<Arg0> man in black shirt	<Arg0> man with tie	<Arg0> man in suit
<Arg1> man in brown shirt	<Arg1> man in black shirt	<Arg1> man in black shirt
<ArgM> across the room		

CONTENT

Application VI:

Generating Visual Spatial Description via Holistic 3D Scene Understanding

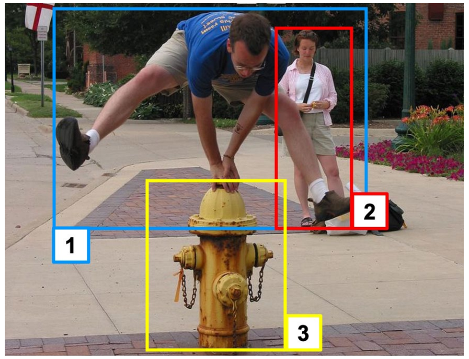
[1] Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, Tat-Seng Chua. Generating Visual Spatial Description via Holistic 3D Scene Understanding. ACL. 2023.

3D-SG-based Visual Spatial Description

Motivation

- Visual Spatial Description (VSD)

Inputs: img, two objects



<‘man’, [2]>
<‘fire hydrant’, [3]>

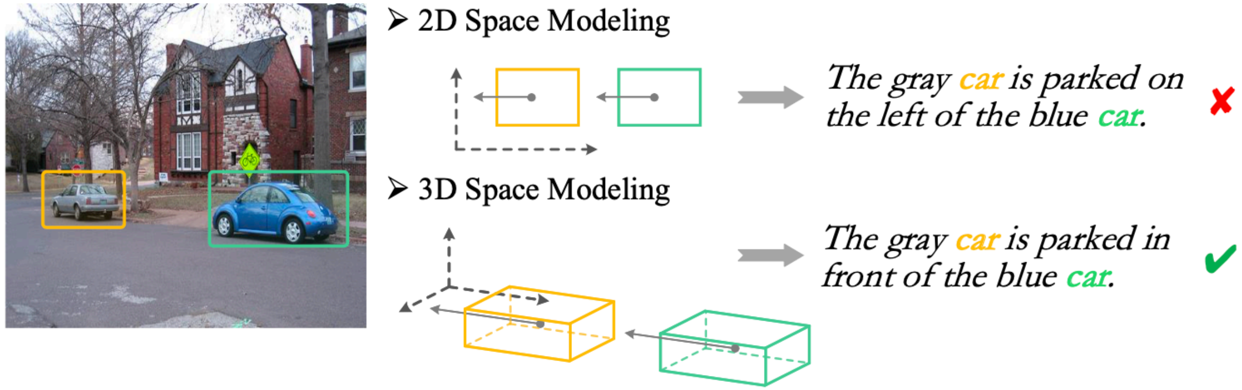
Output: spatial description

*The man in white is **standing behind** the yellow fire hydrant*

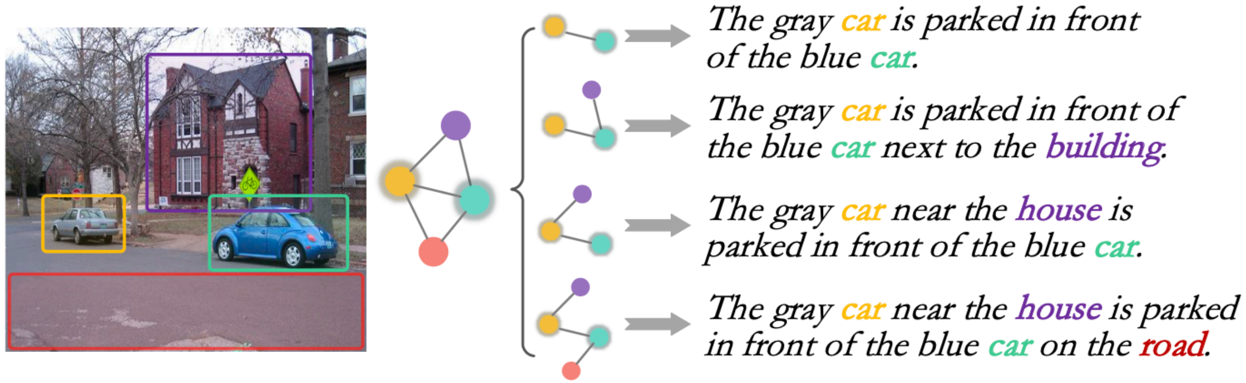
3D-SG-based Visual Spatial Description

Motivation

- Existing issues
  - 2D modeling is NOT enough
    - *Perspective Illusion*
    - *Overlap*
  - Relation descriptions NOT diversified enough
    - *Spatial Diversity*



(a) Modeling 3D scene features results in correct spatial understanding

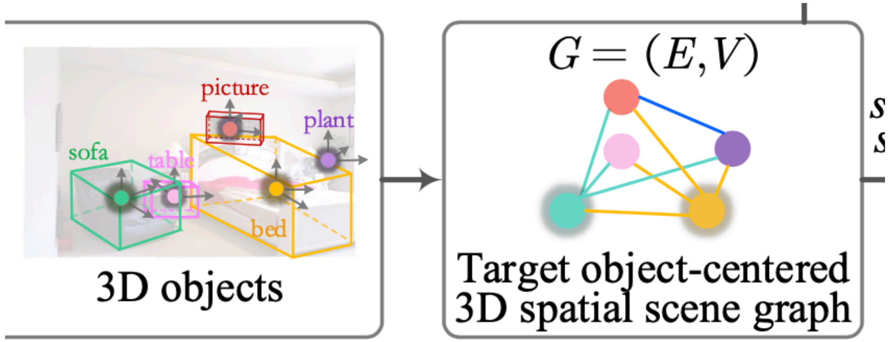


(b) Holistic 3D scene features help generate diversified spatial descriptions

3D-SG-based Visual Spatial Description

Method

- Modeling 3D Scene Graph
  - 3D Scene Feature Extracting
    - *Parsing with an off-the-shelf model*
  - Graph Modeling
    - *Target Object-Centered 3D Spatial Scene Graph (GO3D-S2G)*
    - *Object-Centered GCN (OCGCN)*



[1] Yinyu Nie, etc. Total3dunderstanding: Joint layout, object pose andmesh reconstruction for indoor scenes from a single-image. CVPR. 2020.

3D-SG-based Visual Spatial Description

Method

➢ Framework

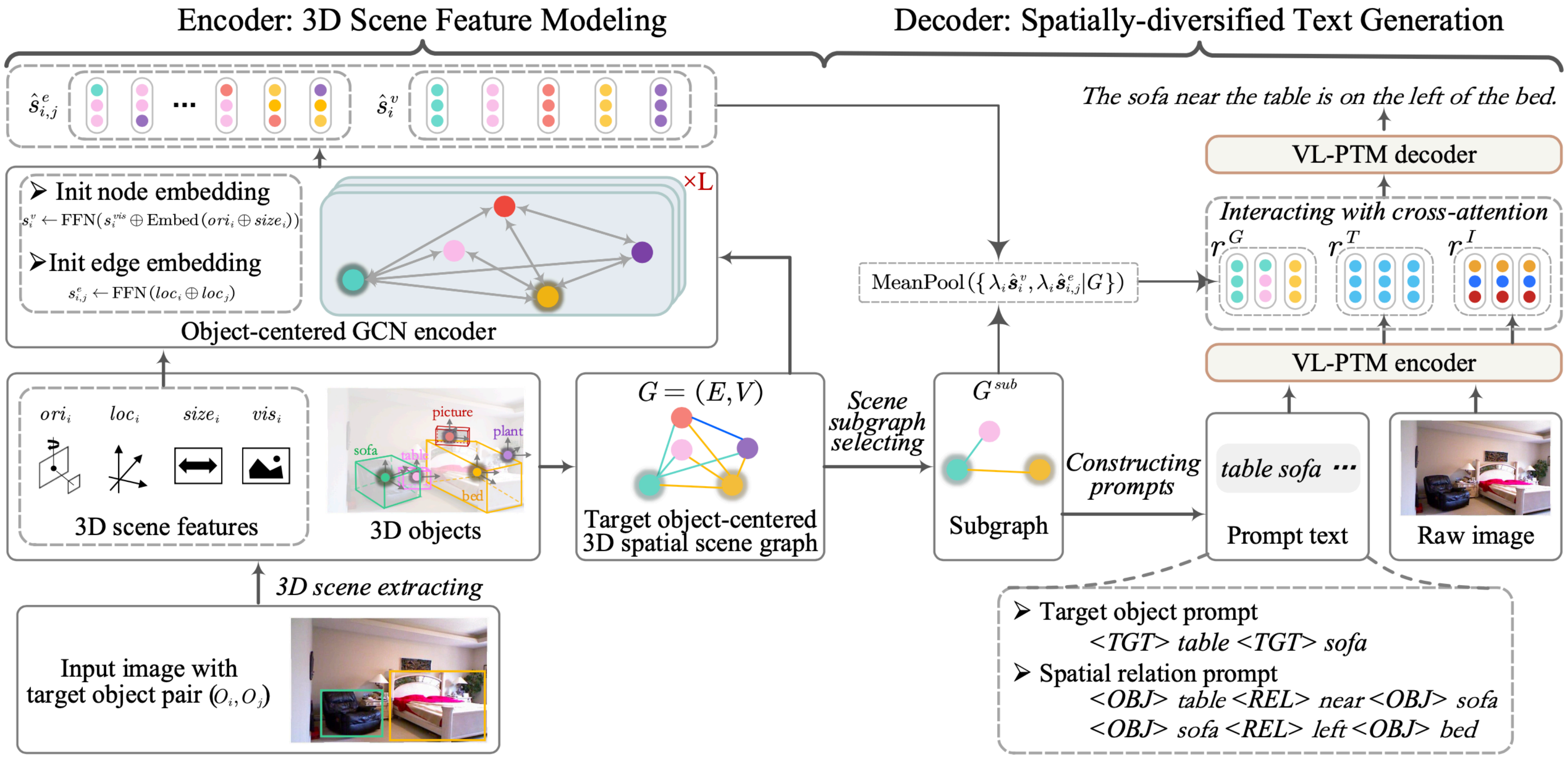
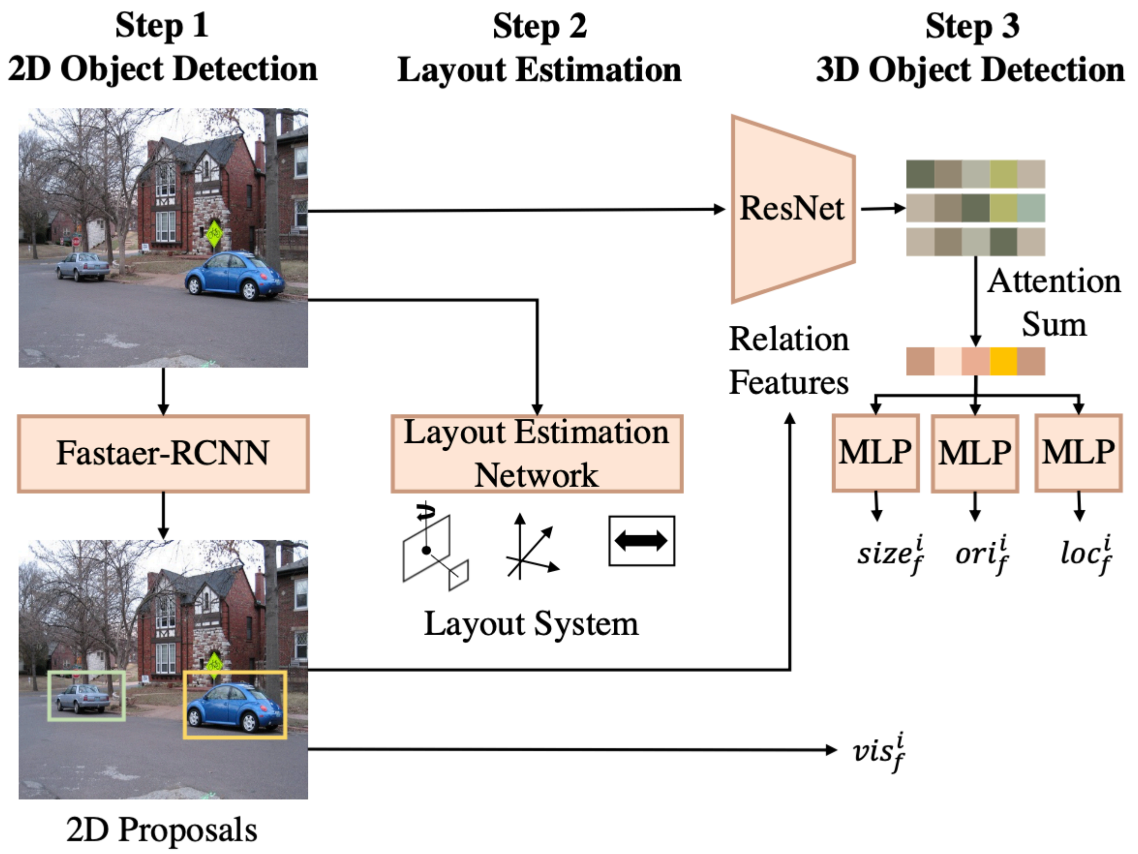


Figure 2: The overview of our proposed framework.

Method

➢ Model Details: 3D Scene Extraction

vis_i	The flatted ROI feature of object i.
$size_i$	The length, width, height of object i.
loc_i	The relative centroid coordinates of object i.
ori_i	The rotation value of three degrees of freedom of object i.



3D-SG-based Visual Spatial Description

Method

➢ Model Details: Graph Creating

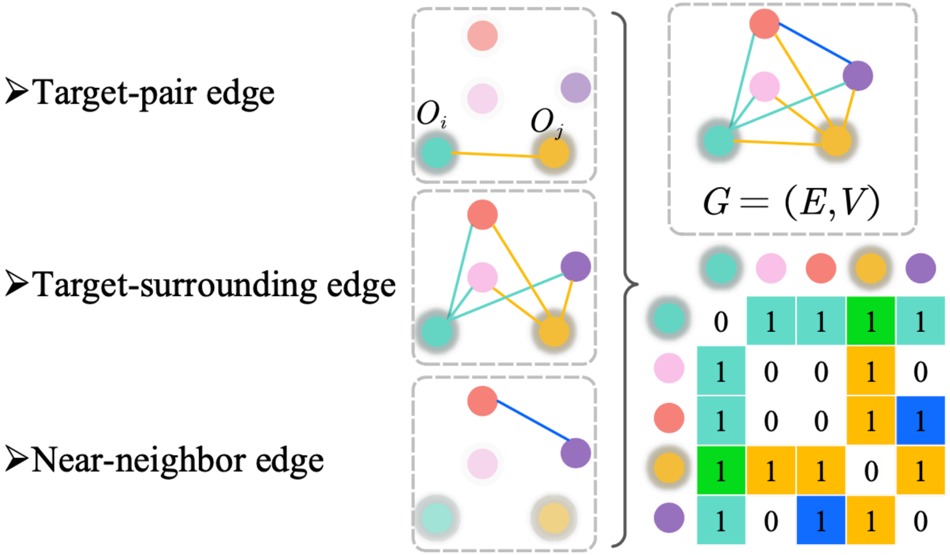


Figure 3: Three types of edges of GO3D-S^{2}G.

Algorithm 1: GO3D-S^{2}G Creating

Input: max object number N,
two target objects index o_1, o_2,
confidency of each object f,
centroid of each object C,
distance threshold d,
noise confidency threshold p

Output: adjacency matrix $\boldsymbol{A}^{N \times N}$

initialization: $\boldsymbol{A} = \mathbf{0}$.

`// target object edges`

$\boldsymbol{A}[o_1,:] = 1$, $\boldsymbol{A}[:, o_1] = 1$,

$\boldsymbol{A}[o_2,:] = 1$, $\boldsymbol{A}[:, o_2] = 1$,

`// add special edges`

for *i in* N **do**
  for *j in* N **do**
    dist = $||C_i - C_j||$
    if *dist* > d **then**
      $\boldsymbol{A}_{ij} = 1$
    end
  end
end

`// remove noise objects`

for *i in* N **do**
  if $f_i < p$ *and* o_i *is not target object* **then**
    $\boldsymbol{A}$[i,:] = 0, $\boldsymbol{A}$[:,i] = 0
  end
end

Method

- Model Details: Scene Subgraph Selecting mechanism (S3)

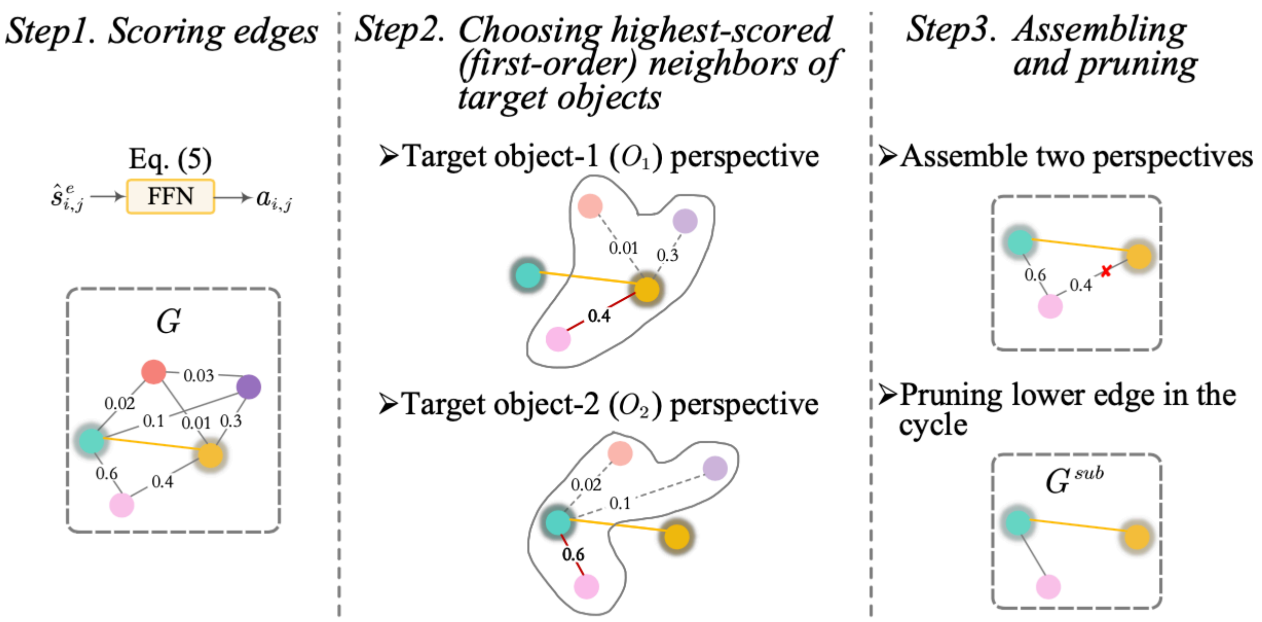
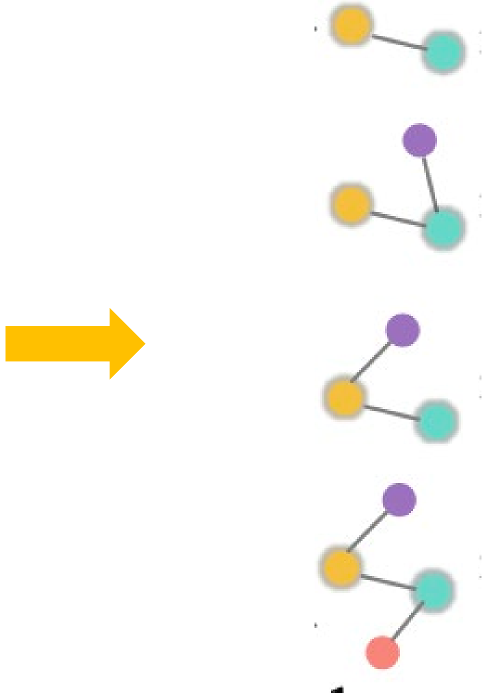


Figure 4: Scene subgraph selecting mechanism.



Method

➢ Model Details: Prompt Learning for LM Decoding

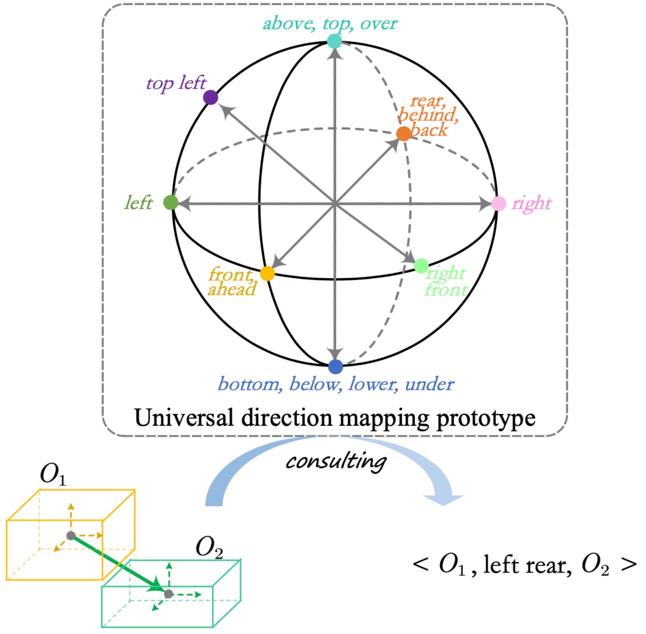


Figure 5: The prototype of direction-term mapping.

prompt texts, 1) *target object prompt*, e.g.,

```
<TGT> table <TGT> sofa
```

and 2) *spatial relation prompt*, e.g.,

```
<OBJ> table <REL> near <OBJ> sofa
<OBJ> sofa <REL> left <OBJ> bed
```

Pre-definitions	
Subject centroid: x_s, y_s, z_s, Object centroid: x_o, y_o, z_o	
coordinate system: x-toward, y-up, z-right	
$x, y, z \in [0, 1]$	
$d_x = \|x_s - x_o\|, d_y = \|y_s - y_o\|, d_z = \|z_s - z_o\|$	
Rule	**Direction Term**
Front: $(d_x > d_y$ and $d_z, d_x > 0.2, x_s > x_o)$	
$d_y, d_z \leq 0.2$	"front"
$d_y > 0.2, y_s > y_o, d_z \leq 0.2$	"front up"
$d_y > 0.2, y_s < y_o, d_z \leq 0.2$	"front down"
$d_z > 0.2, z_s > z_o, d_y \leq 0.2$	"front right"
$d_z > 0.2, z_s < z_o, d_y \leq 0.2$	"front left"
$d_y, d_z > 0.2, y_s > y_o, z_s > z_o$	"front up right"
$d_y, d_z > 0.2, y_s > y_o, z_s < z_o$	"front up left"
$d_y, d_z > 0.2, y_s < y_o, z_s > z_o$	"front down right"
$d_y, d_z > 0.2, y_s < y_o, z_s < z_o$	"front down left"

Back: $(d_x > d_y$ and $d_z, d_x > 0.2, x_s < x_o)$	
$d_x > 0.2, d_y, d_z \leq 0.2$	"back"
Others are similar to front	
Up: $(d_y > d_x$ and $d_z, d_y > 0.2, y_s > y_o)$	
Others are similar to front	
Down: $(d_y > d_x$ and $d_z, d_y > 0.2, y_s < y_o)$	
Others are similar to front	
Right: $(d_z > d_x$ and $d_y, d_z > 0.2, z_s > z_o)$	
Others are similar to front	
Left: $(d_z > d_x$ and $d_y, d_z > 0.2, z_s < z_o)$	
Others are similar to front	
$(d_x, d_y, d_z \leq 0.2)$	"next to"

Table 7: Direction term mapping rules.

3D-SG-based Visual Spatial Description

Experiment

➢ Main Results

	VSD-v1					VSD-v2				
	BLEU-4	**METEOR**	**ROUGE**	**CIDEr**	**SPICE**	**BLEU-4**	**METEOR**	**ROUGE**	**CIDEr**	**SPICE**
• **VL-PTMs**										
Oscar	37.17	35.06	66.47	427.21	67.41	20.90	23.83	50.96	221.61	40.12
VL-Bart	52.71	41.96	77.57	471.21	67.83	20.78	22.83	48.49	213.26	40.04
VL-T5	52.58	41.94	77.63	472.24	67.90	21.83	23.26	50.51	225.51	<u>41.86</u>
OFA	53.59	41.74	77.68	469.23	67.03	<u>22.53</u>	<u>24.93</u>	51.27	227.29	41.63
• **VL-PTMs + VSRC** (Zhao et al., 2022)										
VLBart-ppl	53.49	42.14	77.79	474.34	67.97	21.44	23.08	50.80	226.52	40.16
VLT5-ppl	53.71	42.56	78.33	480.32	68.72	21.79	23.49	51.49	231.70	41.04
VLBart-e2e	53.60	42.45	78.15	476.47	68.18	21.71	23.41	51.22	228.18	40.79
VLT5-e2e	<u>54.31</u>	<u>42.63</u>	<u>78.38</u>	<u>481.13</u>	<u>68.74</u>	22.47	23.50	<u>51.52</u>	<u>231.70</u>	41.07
• **VL-PTMs + 3D scene features**										
3DVSD (Ours)	**56.85** (+2.54)	**43.25** (+0.62)	**79.38** (+1.00)	**483.05** (+1.92)	**68.76** (+0.02)	**26.40** (+3.87)	**26.87** (+1.94)	**55.76** (+4.24)	**272.93** (+41.23)	**46.97** (+5.11)

Table 2: Main results on two datasets. Bold numbers are the best, and underlined ones are the second best.

3D-SG-based Visual Spatial Description

Experiment

- 2D v.s 3D modeling

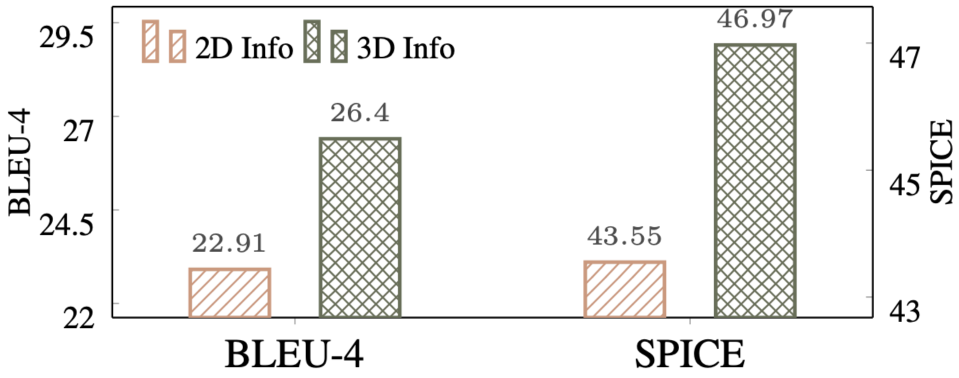


Figure 6: Comparison of 2D and 3D method on VSDv2.

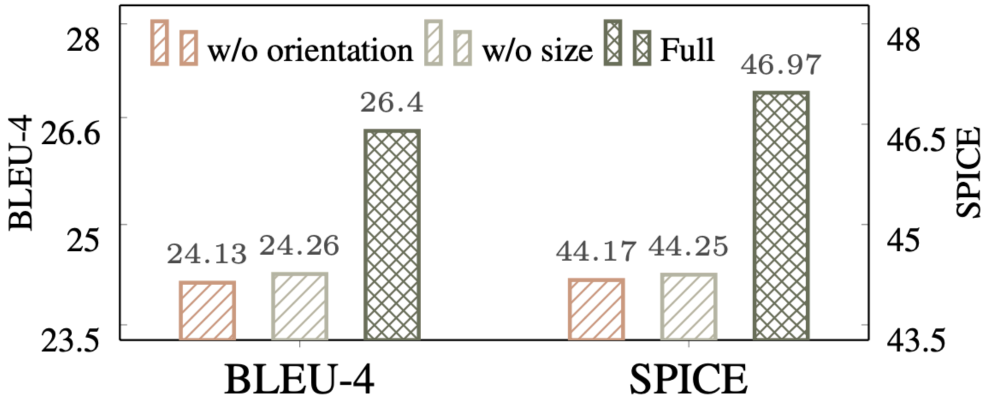


Figure 7: Ablation results of 3D features on VSDv2.

Experiment

➢ Case Study

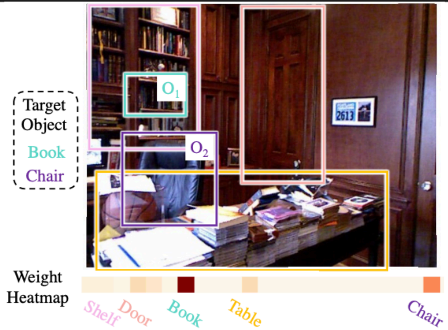
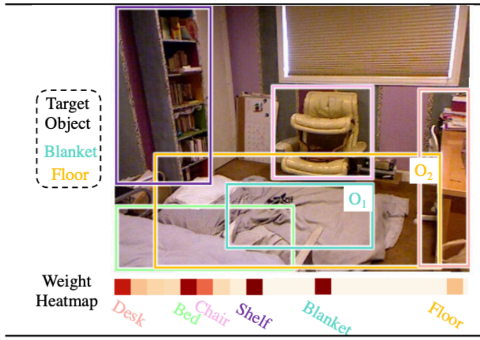
Input	With Beam Search	With Scene Subgraph Sampling
	• **VLT5-e2e:** The books are on the chair. There are some books above the chair. Some books are on the black chair. • **3DVSD:** The book is behind the chair. Some books are behind the chair. There are some books behind the black chair.	• **3DVSD:** The books on the shelf are behind the chair. Some books are on the shelf behind the chair. The books are behind the chair next to the table. The books on the shelf are behind the chair near the door. The books on the shelf are behind the chair next to the table.
	• **VLT5-e2e:** The blanket is near the floor. The gray blanket is under the floor. There is a white blanket on the floor. • **3DVSD:** The blanket is on the floor. The white blanket is on the floor. The grey blanket is on the floor.	• **3DVSD:** The gray blanket is on the floor. The blanket on the floor is in front of the chair. The blanket on the floor is on the right of the bed. The blanket on the floor is in front of the shelf. The blanket is on the floor next to the desk.

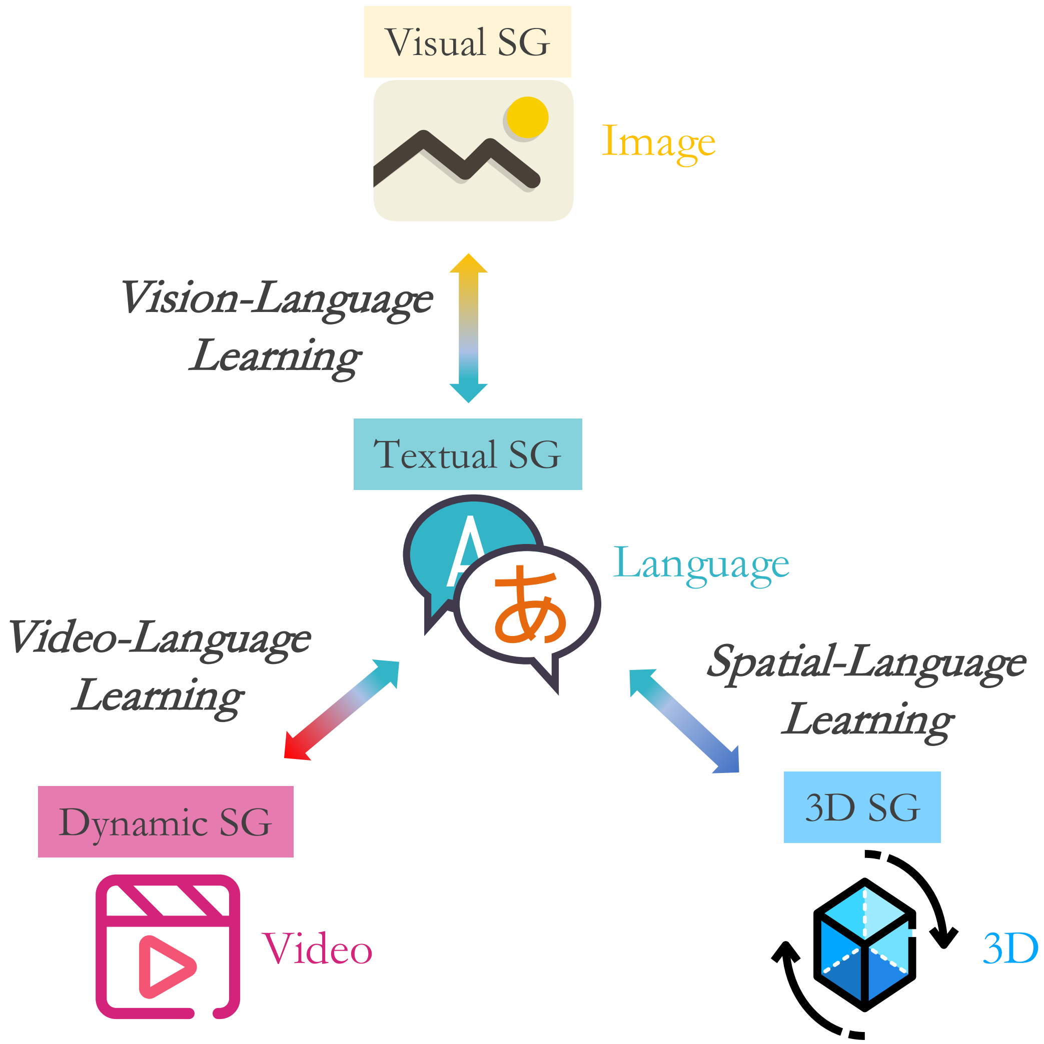
Figure 9: Qualitative results of generated descriptions with beam search decoding and S^3 mechanism, respectively.

CONTENT

Outlook of Future Directions

Summary

- Vision&Language Scene Graph Modeling
- Video Scene Graph Modeling
- 3D Scene Graph Modeling



Outlook of Future Directions

What Next?

- Applying Scene Graph Representations into More Scenarios and Applications
  - Image/Video Retrieval
  - Image/Video Editing
  - Image/Video Generation
  - Video Moment Localization
  - …

1. Improving cross-modal alignment:
   more fine-grained vision-text matching
2. Enhancing multimodal fusion:
   semantic-level feature learning
3. More controllable end-task prediction:
   highly structured modal representation

Outlook of Future Directions

What Next?

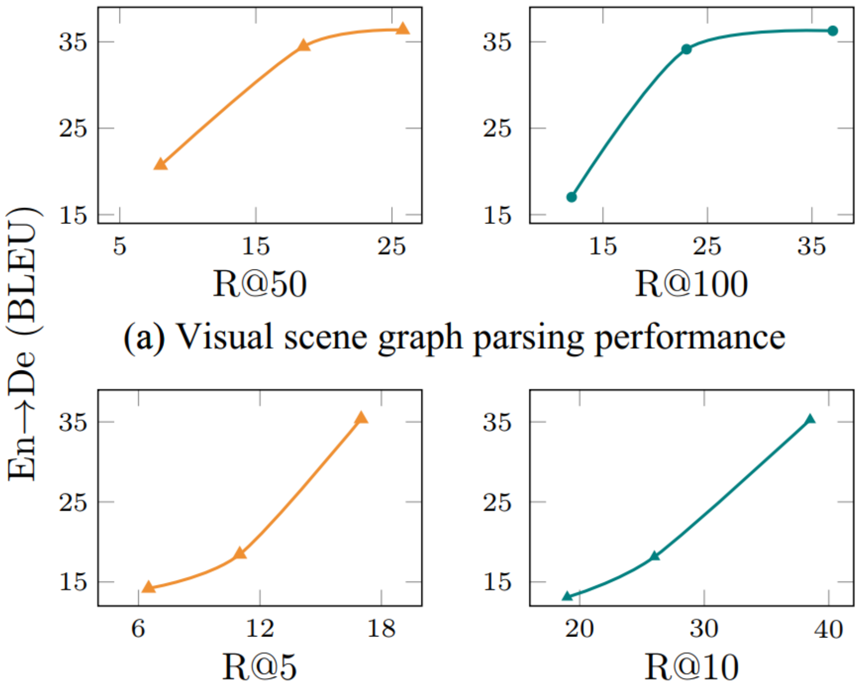
➢ Automatic Learning of Scene Graph Representations

- Low-quality SG annotations decrease the efficacy of the SG features for end tasks.

- How about: Inducing the SG structure along with the end task? Such that the automatically generated SG structures are most coincident with task need.

Latent Structure Induction

Grammar Induction



(a) Visual scene graph parsing performance

(b) Language scene graph parsing performance

Outlook of Future Directions

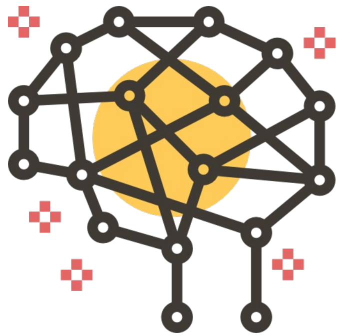
What Next?

- Constructing Semantically Universal Scene Graph (**USG**)
  - Text: *abstract semantics*
  - Image: *detailed semantics*
  - Video: *temporal dynamics*
  - Sound: *vocal attributes*
  - 3D: *depth features*
  - …

World

- Modality-agnostic
- Language-agnostic
- Domain-agnostic

Highly Structured Universal Semantic Representation



World Model

CONTENT

5 Extra delivery

Universal Structured NLP (XNLP) Demo

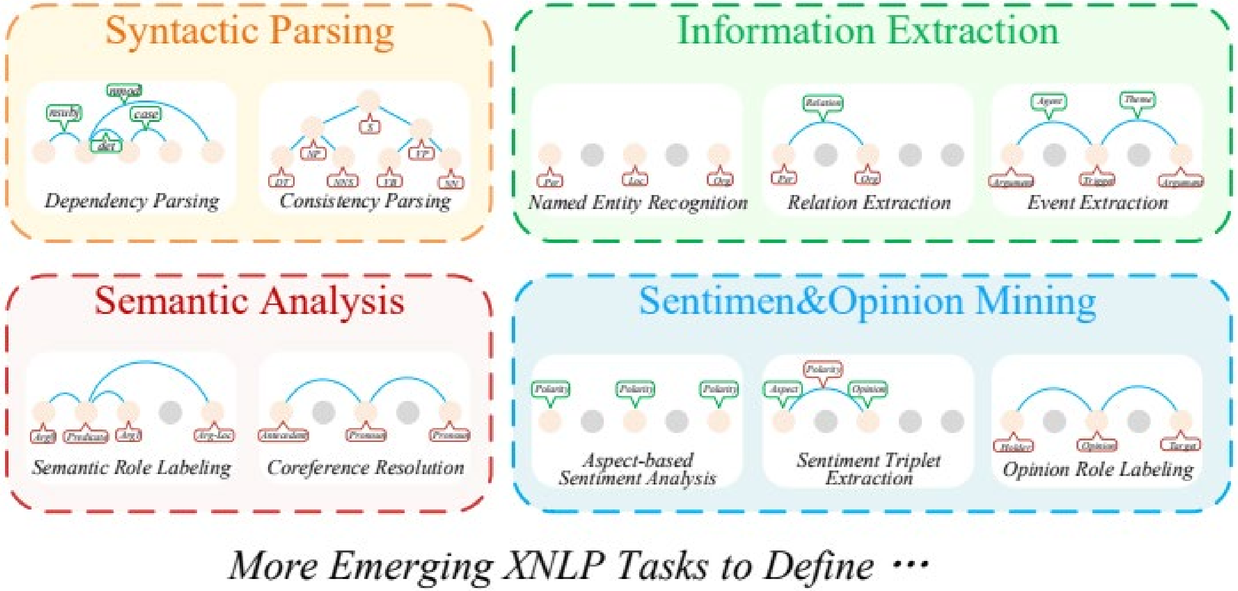
XNLP: An Interactive Demonstration System for Universal Structured NLP

https://xnlp.haofei.vip/

[1] Hao Fei, Meishan Zhang, Min Zhang, Tat-Seng Chua. XNLP: An Interactive Demonstration System for Universal Structured NLP. 2023.

Motivation

- **Structured Natural Language Processing (XNLP)**
  - Many NLP tasks can be reduced into structural predictions
    - 1) textual spans
    - 2) relations between spans



XNLP Demo

Motivation

➢ **Universal XNLP**

- Unified Sentiment Analysis
- Universal Information Extraction

□ a comprehensive and effective approach for unifying all XNLP tasks is not fully established.

➢ **Unification with LLM**

?

✓ One model for all

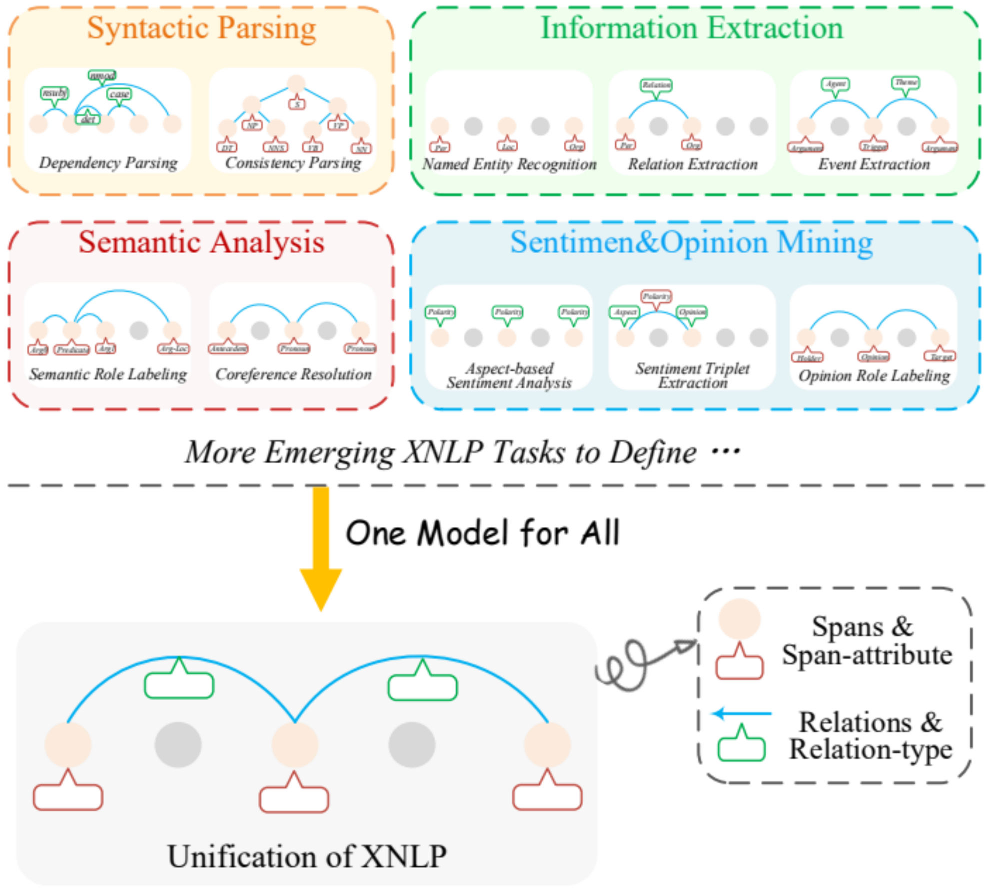
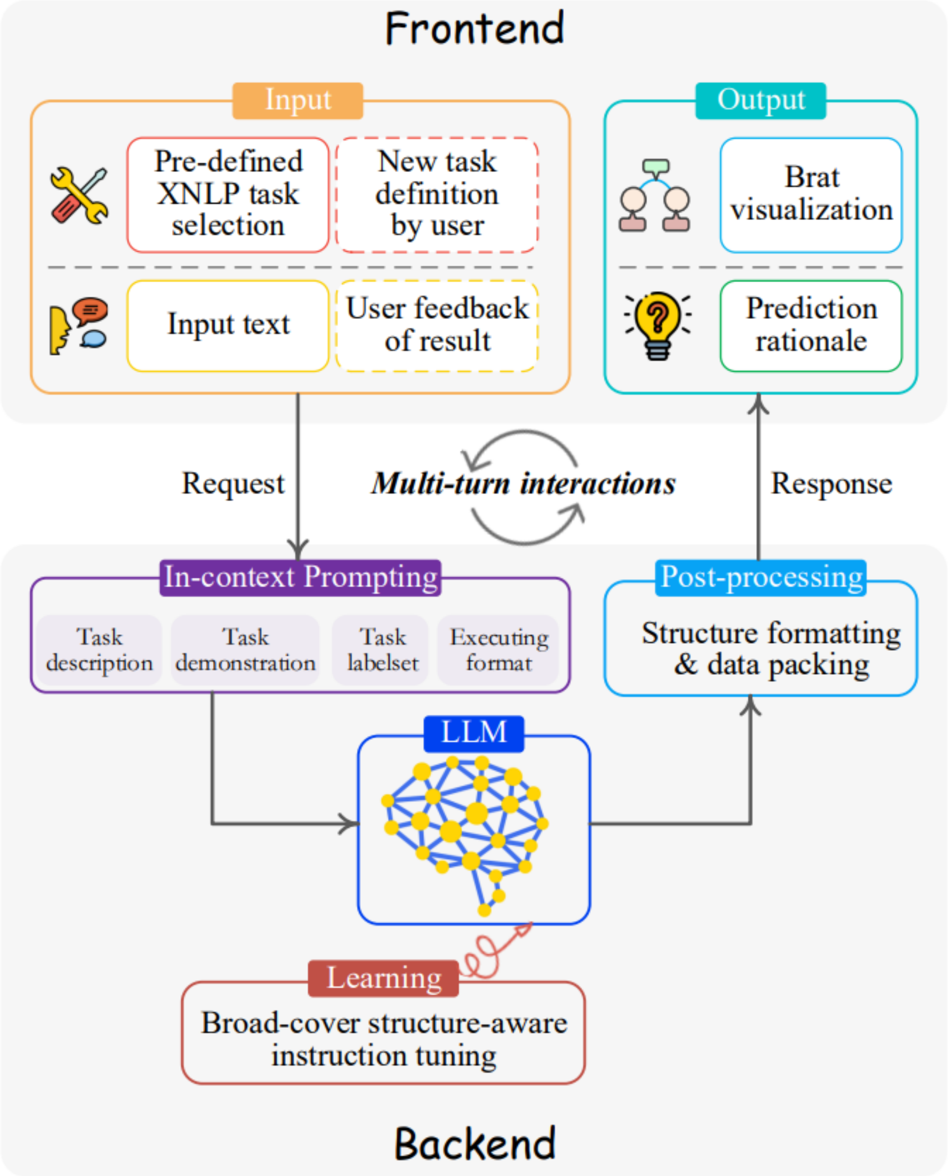


Figure 1: Illustration of the Structured NLP (XNLP) tasks, and the unification of XNLP by decomposing into the predictions of spans and relations.

XNLP Demo

Demo System

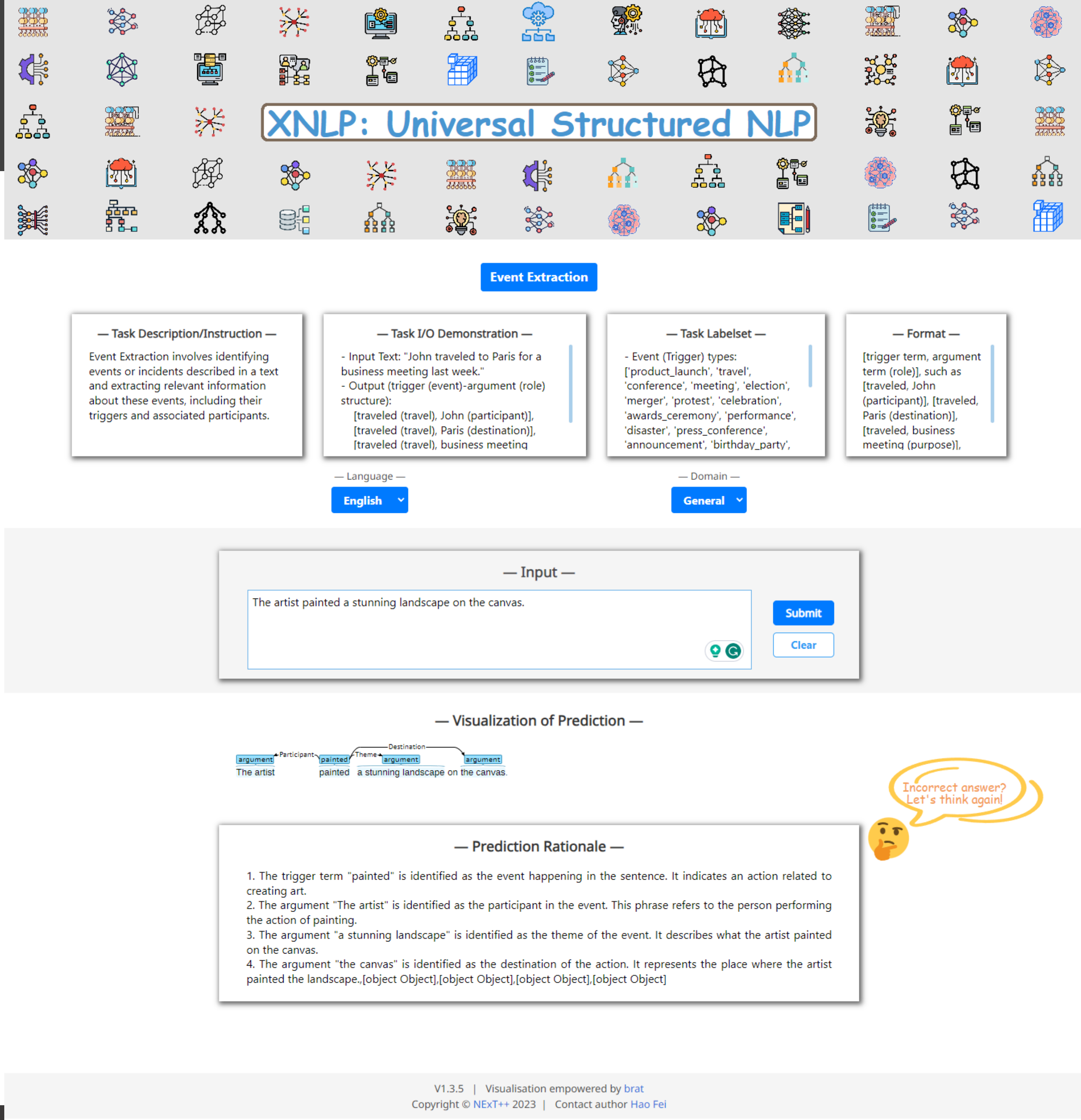
- System Design



XNLP Demo

Demo System

- Screenshot

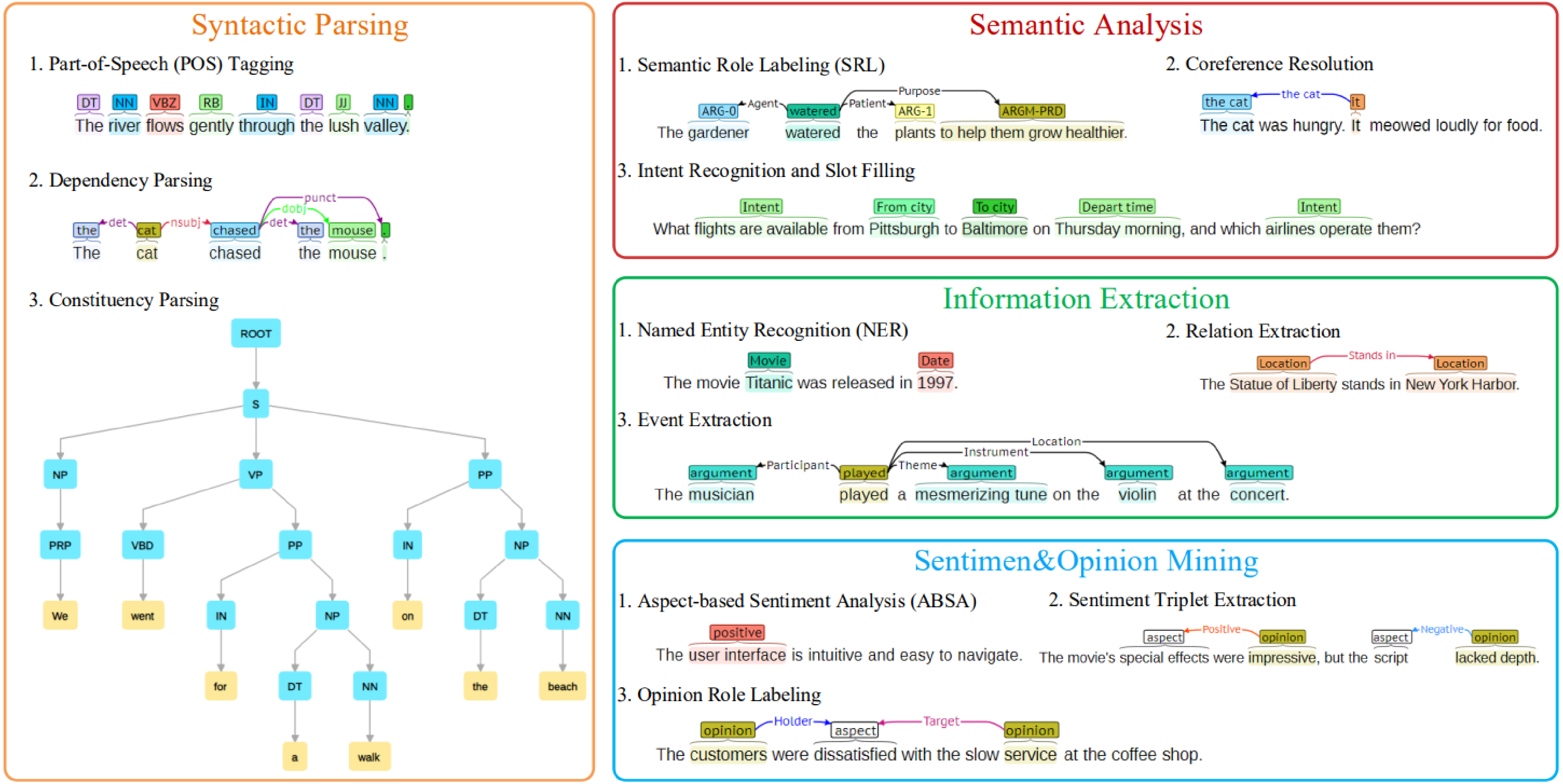


XNLP Demo

NExT++

Demo System

https://xnlp.haofei.vip/



Thanks

Q&A