LREC-COLING 2024

TARN-VIST: Topic-Aware Reinforcement Network for Visual Storytelling

Weiran Chen, Xin Li, Jiaqi Su, Guiqian Zhu, Ying Li, Yi Ji, Chunping Liu

School of Computer Science and Technology, Soochow University, Suzhou, China

Introduction - Task Definition

Visual Storytelling: generate a story for an ordered image sequence automatically

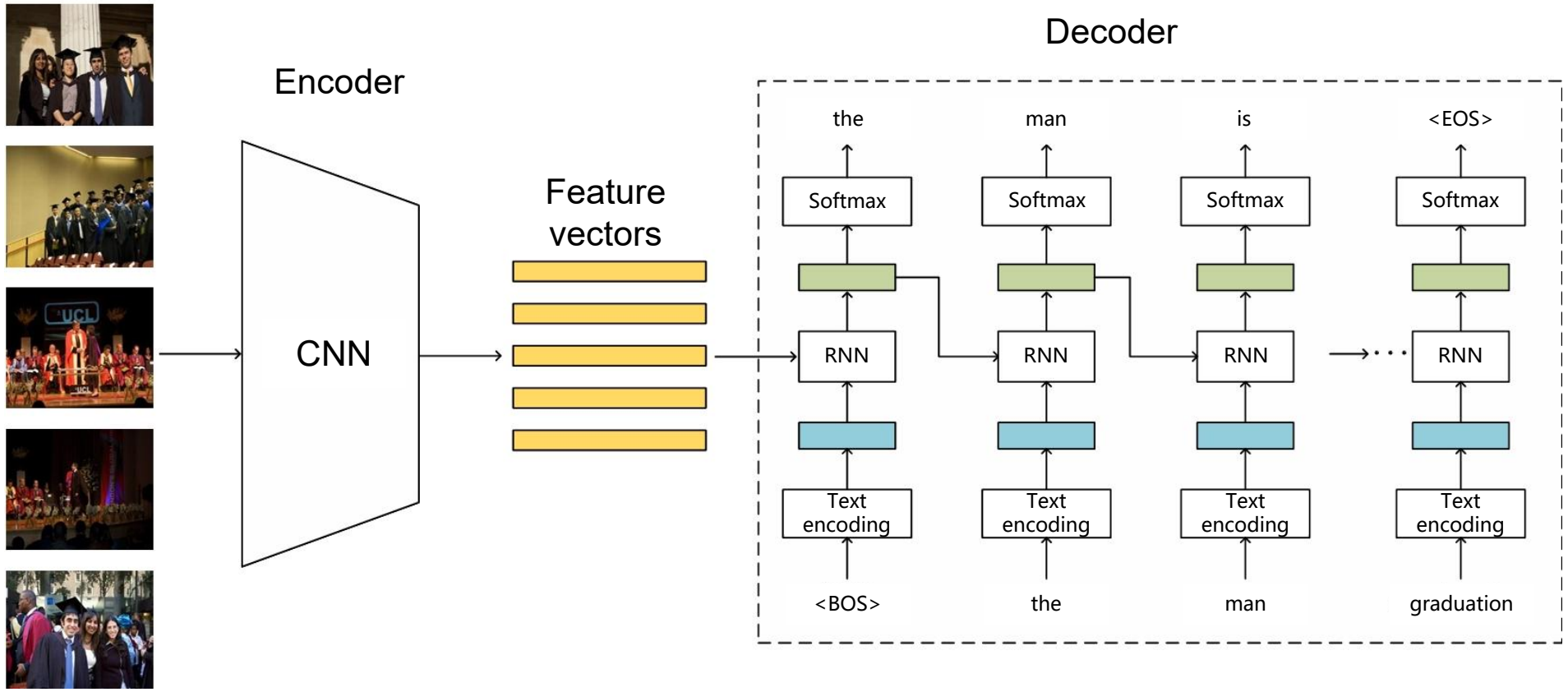
A discus got stuck up on the roof. Why not try getting it down with a soccer ball? Up the soccer ball goes. It didn't work so we tried a volleyball. Now the discus, soccer ball, and volleyball are all stuck on the roof.

The dog was ready to go. He had a great time on the hike, and was very happy to be in the field. His mum was so proud of him. It was a beautiful day for him.

Introduction - Approach Summary

End-to-end based Approach

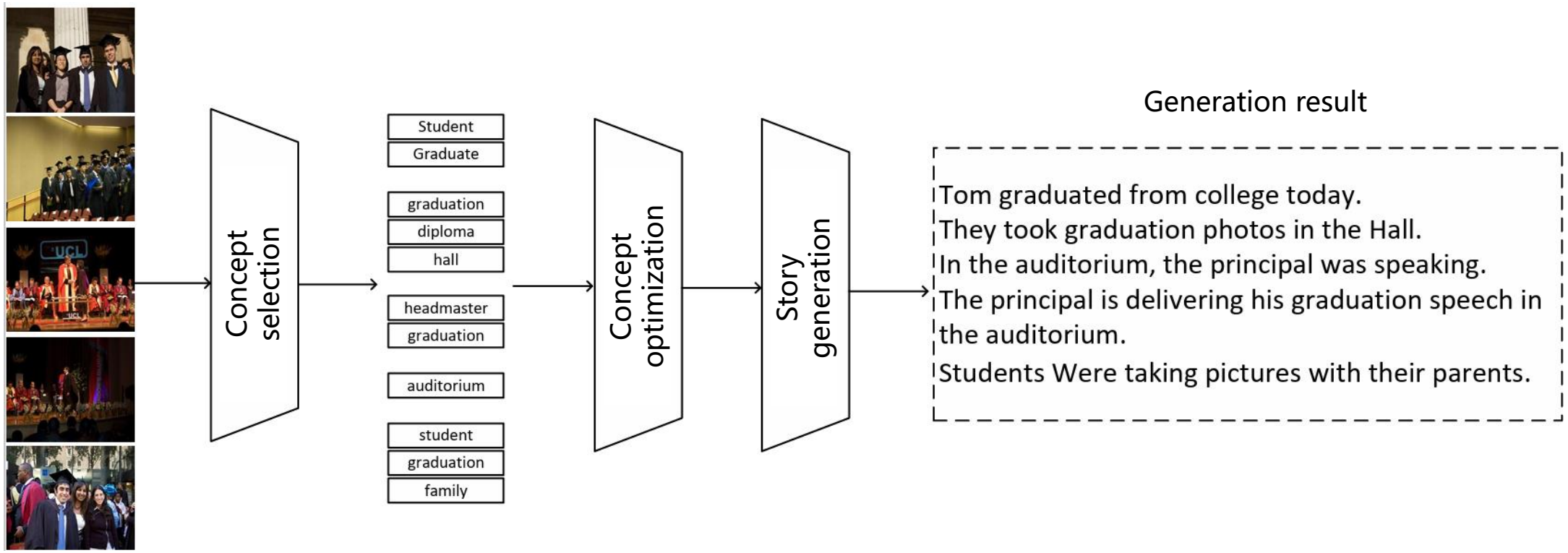
- Utilize a CNN as an encoder to extract image features and overall image-stream features
- Feed the feature vectors into a RNN to construct the story



Introduction - Approach Summary

Multi-stage based Approach

- Separate the generation process into multiple steps
- The output of the previous step is often leveraged as the input of the subsequent step



Our work

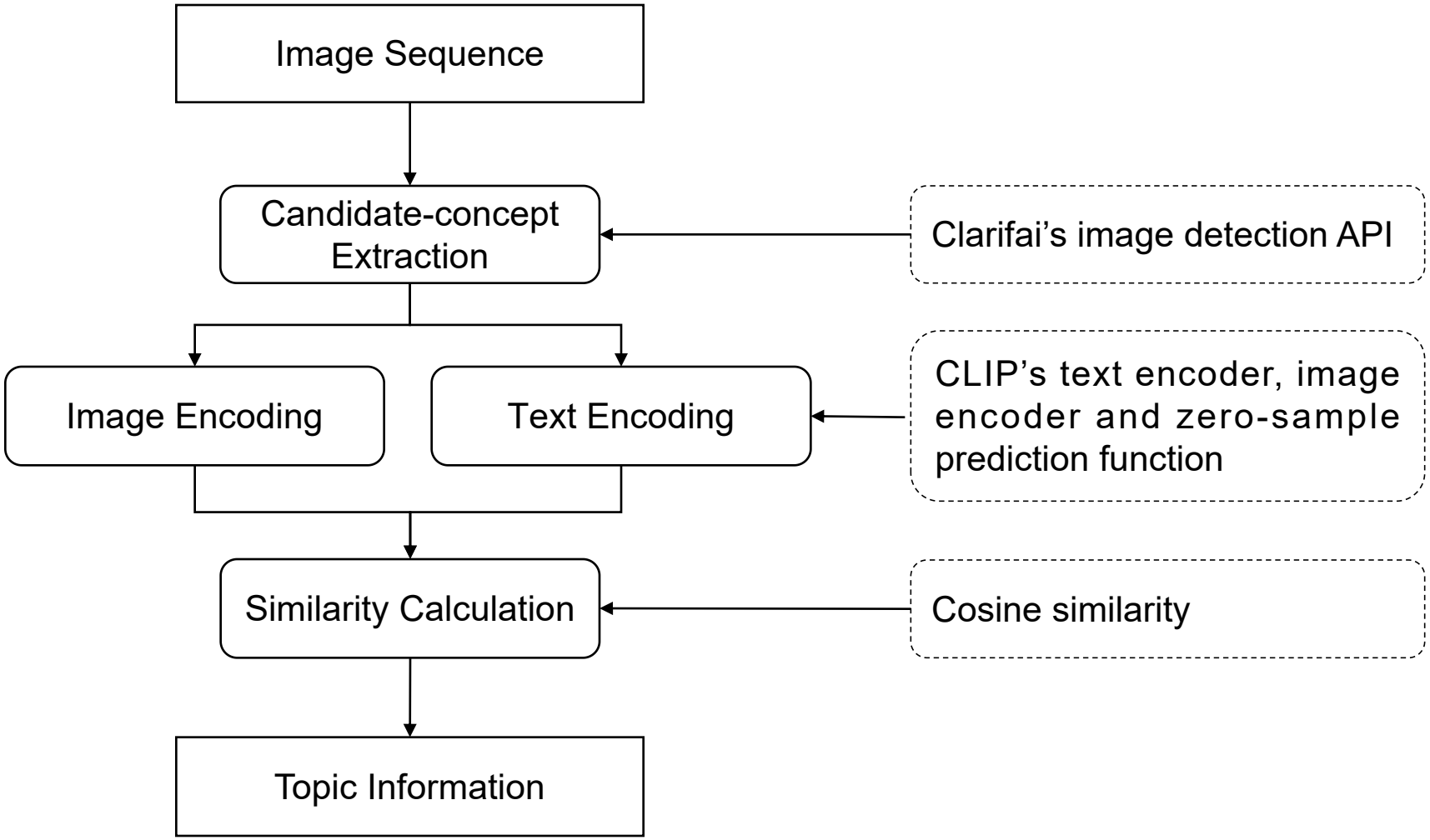
Motivations

- Few models take the latent topic information of the generated story into account
- Previous work does not consider extracting topic words from the visual aspect

Our Solution: Topic Aware Reinforcement Network for Visual Storytelling (TARN-VIST)

- Use CLIP and RAKE to extract topic information from both visual and linguistic perspectives
- Design reinforcement learning rewards for topic consistency based on the topic information and cosine similarity
- Experimental results show that our method outperforms most of the leading models on multiple evaluation metrics.

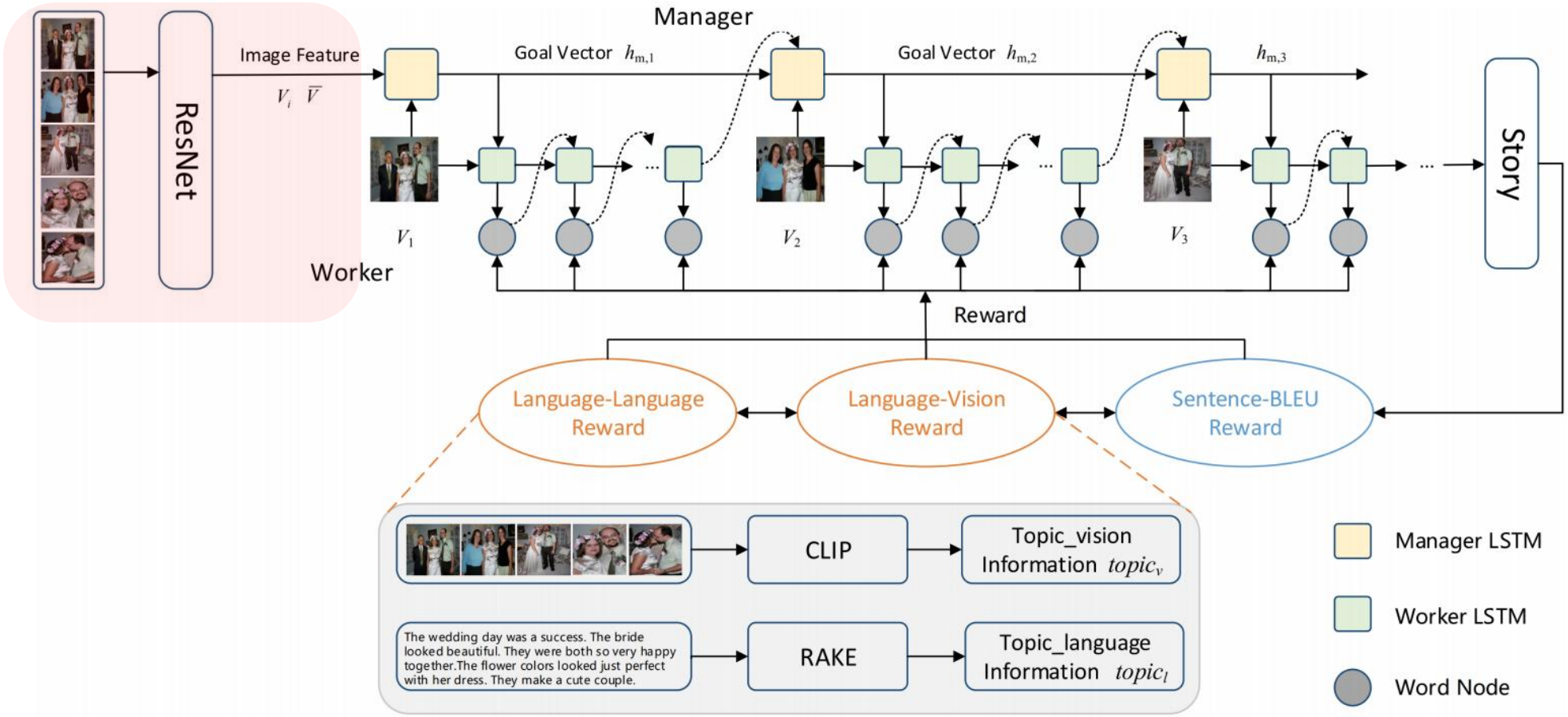
TARN-VIST - Topic Information Extraction



TARN-VIST - Topic Information Extraction

Images					
Topic Information	Topic_vision: Seashore; Topic_language: Good Day;				
Images					
Topic Information	Topic_vision: Birthday; Topic_language: Party;				
Images					
Topic Information	Topic_vision: Parade; Topic_language: Parade;				

TARN-VIST - Topic Aware Storytelling Model



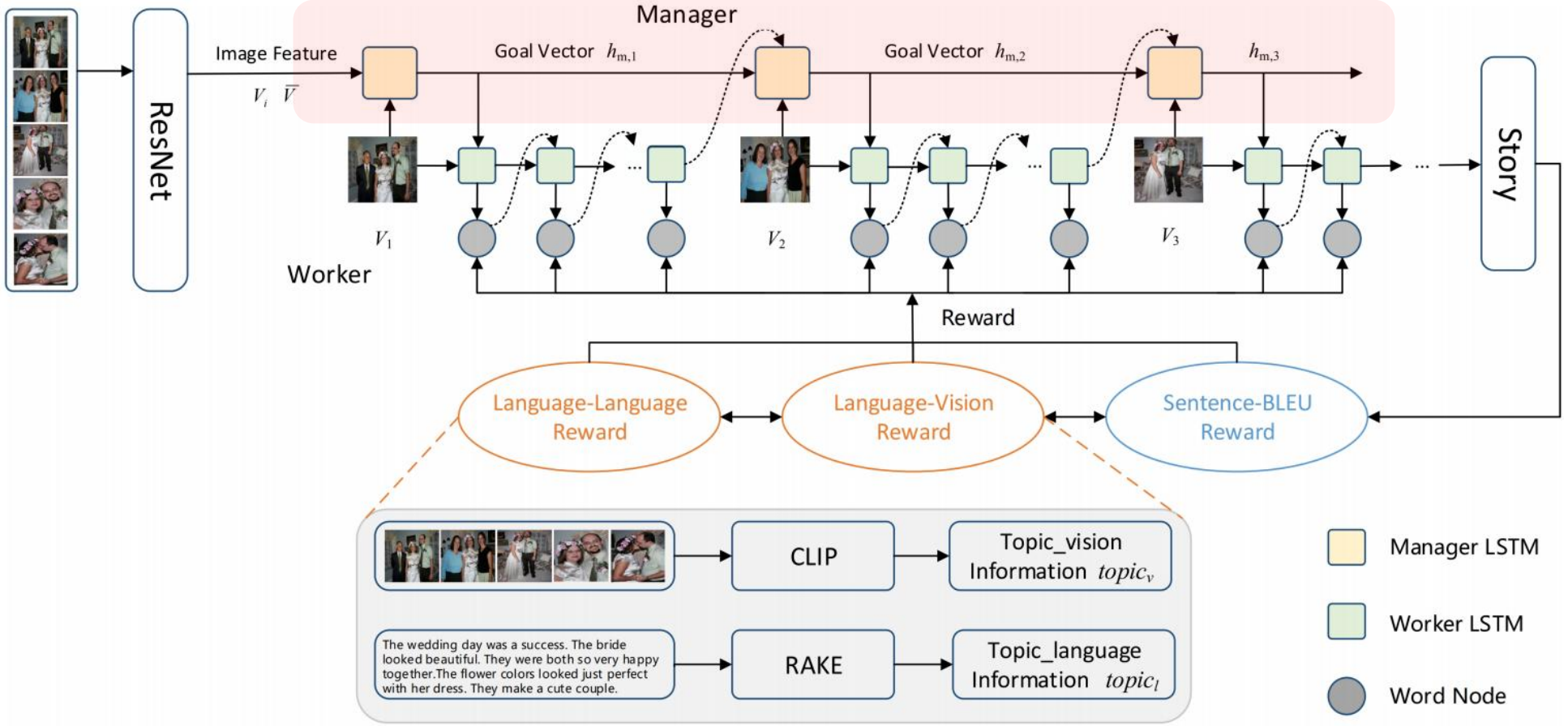
$$v_i = \text{Resnet}(I_i)$$

$$\overline{V} = \frac{1}{n}\sum_i v_i$$

v_i : the individual feature information of image i

$\overline{V}$: overall feature information of the image sequence I

TARN-VIST - Topic Aware Storytelling Model



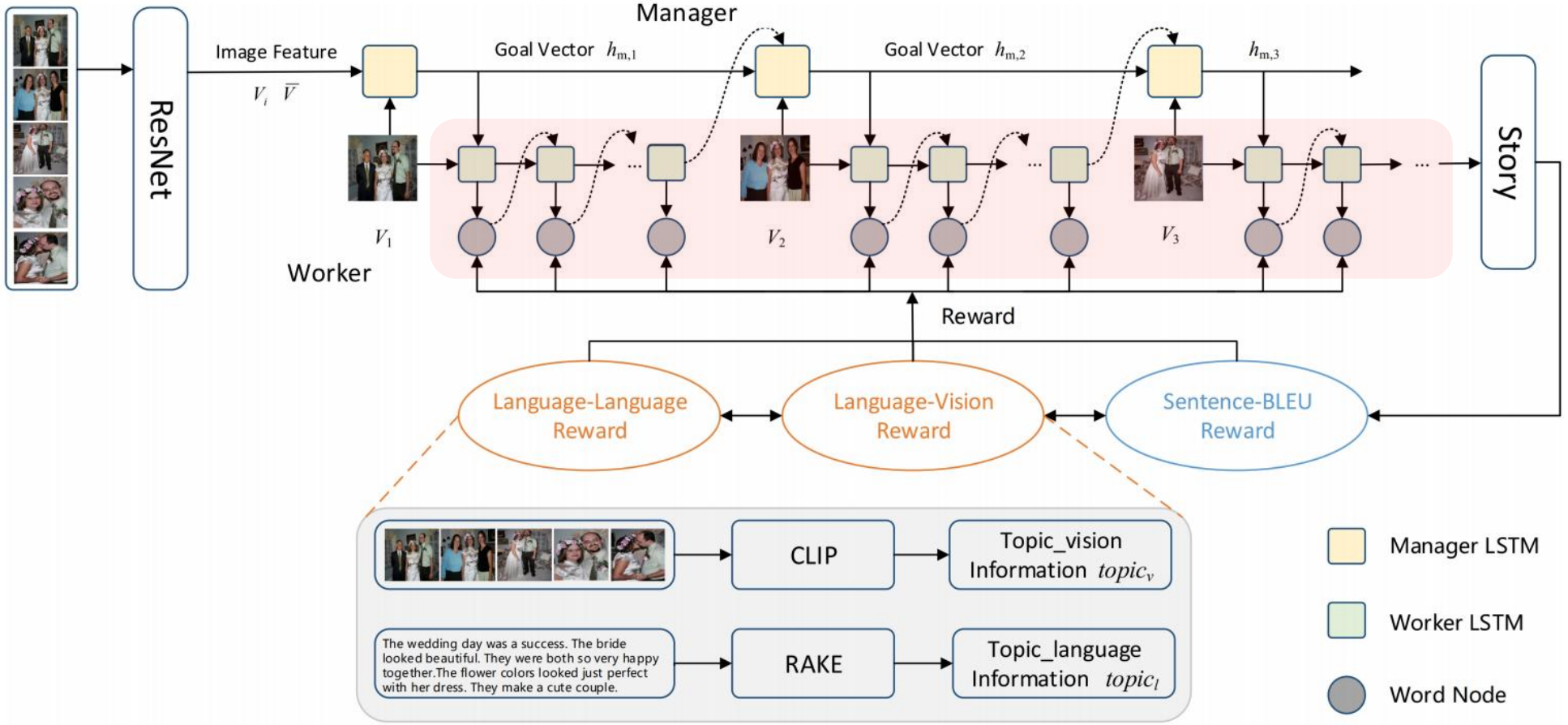
Manager LSTM: Serve as a supervisor to control the overall flow of the story

$$h_{m,i} = \mathrm{LSTM}_M([\overline{V}; v_i; h^t_{w,i-1}], h_{m,i-1})$$

$h^t_{w,i-1}$: sentences generated for the previous image

$h_{m,i}$: the hidden state (goal vector)

TARN-VIST - Topic Aware Storytelling Model



Worker LSTM: Complete the generation of word description based on the goal vector $h_{m,i}$

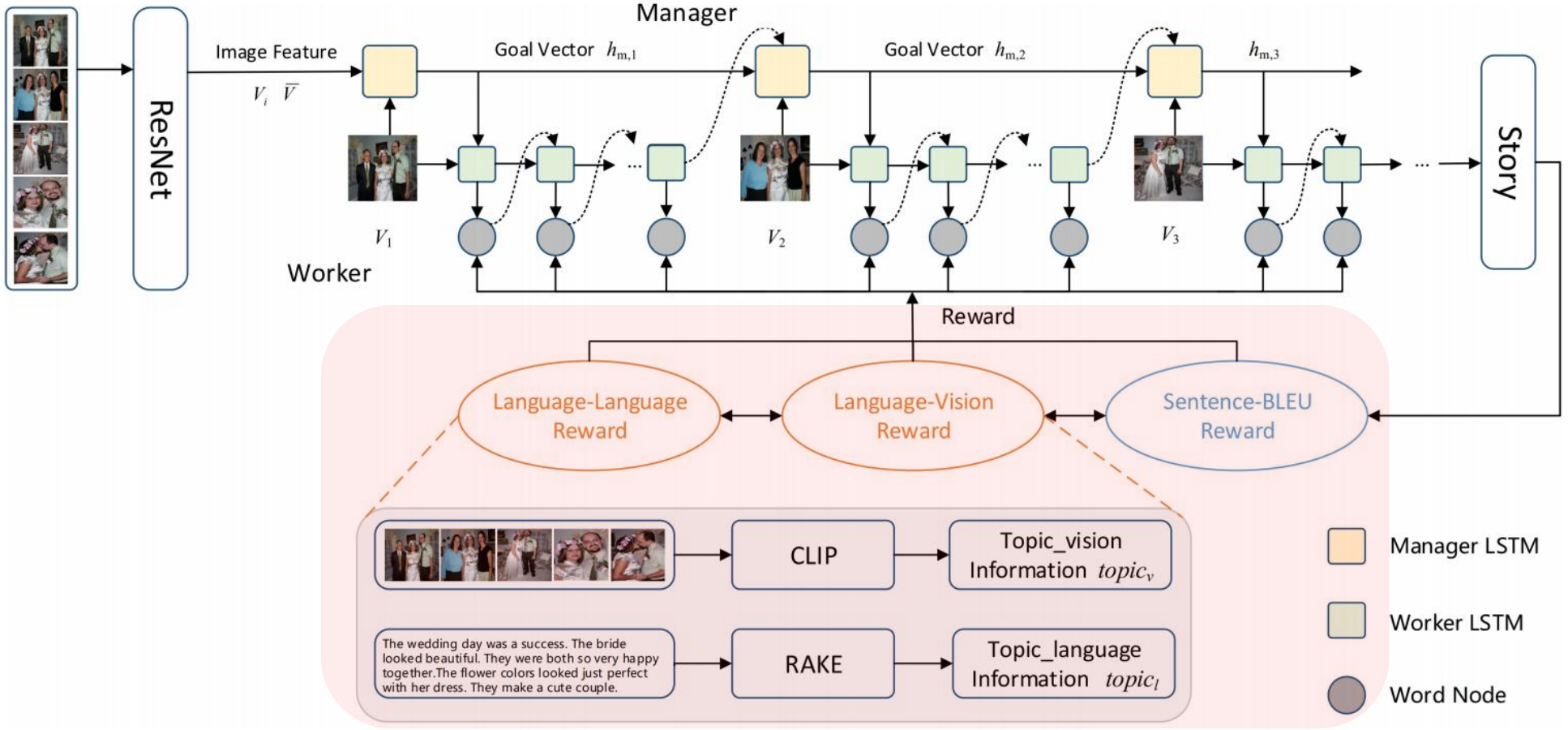
$$h_{w,i}^t = \mathrm{LSTM}_W([v_i; h_{m,i}; e_i^{t-1}], h_{w,i}^{t-1})$$

$$p_\theta(y_i^t \mid y_i^{1:t-1}; v_i, \overline{V}) = \mathrm{softmax}(f(h_{w,i}^t))$$

e_i^{t-1} : the word embedding of previously generated word

$f(\cdot)$: linear layer

TARN-VIST - Topic Aware Storytelling Model



Topic Consistency Rewards Design:

$$r_{topic-cv} = \text{cosine-similarity}(topic_c, topic_v) \qquad r_{bleu} = \text{sentence-bleu}(story_c, story_g)$$

$$r_{topic-cl} = \text{cosine-similarity}(topic_c, topic_l) \qquad r = \lambda \cdot r_{bleu} + \gamma \cdot r_{topic\text{-}cv} + \eta \cdot r_{topic\text{-}cl}$$

Model Training

Reinforcement loss: encourage the model to focus on key aspects via maximizing the reward

$$Loss_{RL}(\beta) = \sum_{Y,V \in D'} E_{y_i \sim \pi_i}[(b - r(y_i)) \log \pi_i]$$

$$r(y_i) = \lambda \cdot r_{bleu}(y_i) + \gamma \cdot r_{topic\text{-}cv}(y_i) + \eta \cdot r_{topic\text{-}cl}(y_i)$$

$$\pi \equiv p_\theta(y_i \mid v_i, \bar{V})$$

Two-stage training strategy

- First stage: Train with maximum likelihood estimation (MLE)
- Second stage: Train with jointly with reinforcement loss and MLE loss

$$Loss_{MLE} = \sum_{Y,V \in D'} \sum_{i=1}^{N} \sum_{t=1}^{T} -\log P(W_{i,t} = Y_{i,t})$$

$$Loss_{mixed} = \omega \cdot Loss_{RL} + (1 - \omega) \cdot Loss_{MLE}$$

Experimental Results - Quantitive Evaluation

Table 1: Quantitive results on the VIST dataset for surface-level-based automatic metrics. For all these metrics, higher score means better performance.

Method	BLEU-1	BLEU-2	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Seq2seq	-	-	3.5	31.4	-	6.8	-
H-Attn-Rank	-	-	-	34.1	29.5	7.5	-
BARNN	-	-	-	33.3	-	-	-
SRT	43.4	21.4	5.2	12.3	-	11.4	-
XE-ss	62.3	38.2	13.7	34.8	29.7	8.7	-
AREL	63.7	39.0	14.0	35.0	29.6	9.5	8.9
HPSR	61.9	37.8	12.2	34.4	**31.2**	8.0	-
HSRL	-	-	12.3	35.2	30.8	10.7	7.5
SGVST	65.1	40.1	14.7	**35.8**	29.9	9.8	-
ReCo-RL	-	-	12.4	33.9	29.9	8.6	8.3
INet	64.4	40.1	14.7	35.6	29.6	11.0	-
IRW	66.7	41.6	**15.4**	35.6	29.6	11.0	-
CKAKS	-	-	12.0	35.4	30.0	10.5	-
LGMT	67.5	41.6	15.1	35.6	29.7	10.0	-
Sentistory	64.8	39.8	14.2	35.3	29.8	9.7	-
TARN-VIST	**69.0**	**43.5**	13.4	**35.8**	29.5	**12.1**	**11.3**

Experimental Results - Quantitive Evaluation

Table 2: Quantitive results on the VIST dataset for semantic understanding evaluation metric. For all these metrics, higher score means better performance.

Method	BERTScore	BARTScore	BLEURT
KE-VIST (No KG)	28.25	17.21	43.63
KE-VIST (With OpenIE)	29.12	17.93	46.85
KE-VIST (With KG)	29.16	18.03	47.54
PR-VIST	27.64	18.09	48.92
TARN-VIST	**30.47**	**18.51**	**49.43**

Experimental Results - Ablation Study

Table 3: Ablation experiment results on different combinations of the reward functions. Note that our basic model is ReCo-RL.

Method	BLEU-4	METEOR	CIDEr	SPICE
Baseline	12.40	33.90	8.60	8.30
Baseline+r_{bleu}	12.82	35.36	11.58	10.84
Baseline+r_{bleu}+$r_{topic\text{-}cv}$	13.10	35.62	11.70	**11.37**
Baseline+r_{bleu}+$r_{topic\text{-}cl}$	13.44	35.80	10.99	11.36
TARN-VIST	**13.46**	**35.88**	**12.07**	11.25

$$r_{topic-cv} = \text{cosine-similarity}\,(topic_c, topic_v)$$

$$r_{topic-cl} = \text{cosine-similarity}\,(topic_c, topic_l)$$

$$r_{bleu} = \text{sentence-bleu}\,(story_c, story_g)$$

Experimental Results - Hyper-parameter Selection

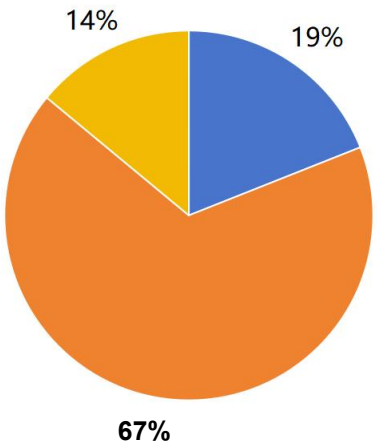
Table 4: Experiment results of TARN-VIST with different γ and η.

γ	η	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
0.3	0.7	68.43	42.48	23.51	12.85	29.14	35.52	11.23	10.76
0.4	0.6	68.38	42.55	23.56	12.88	29.21	35.62	11.69	10.85
0.5	0.5	**69.01**	**43.56**	**24.27**	**13.46**	**29.28**	**35.88**	**12.07**	**11.36**
0.6	0.4	68.58	42.79	23.77	13.09	29.24	35.63	11.91	10.94
0.7	0.3	68.24	42.46	23.57	12.92	29.12	35.53	11.44	10.78
0.8	0.2	68.24	42.58	23.66	12.90	12.90	35.48	11.69	10.77

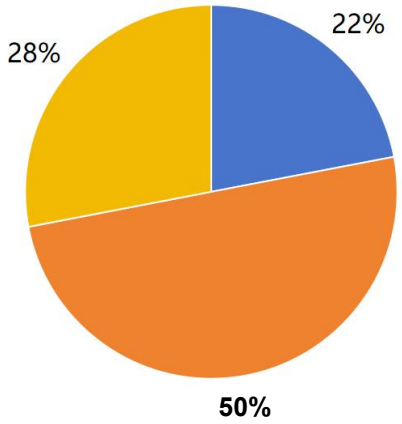
$$r = \lambda \cdot r_{bleu} + \gamma \cdot r_{topic\text{-}cv} + \eta \cdot r_{topic\text{-}cl} \qquad (\lambda = 0.5,\ \gamma + \eta = 1)$$

Experimental Results - Human Evaluation

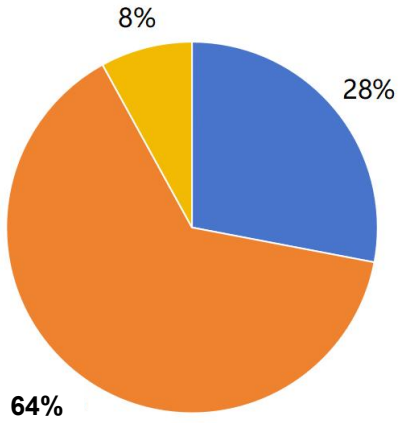
Relevance

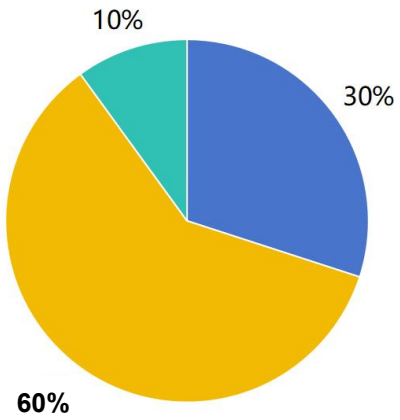


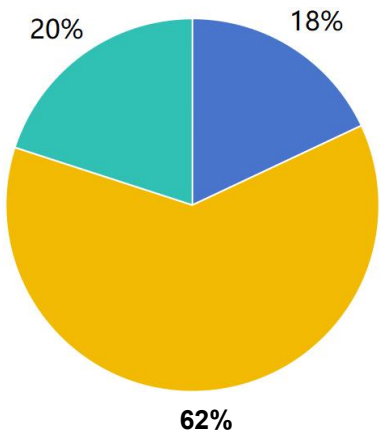
Coherence

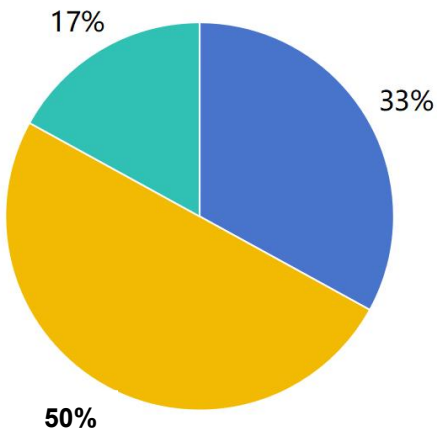


Information Richness









Experimental Results - Qualitative Evaluation

Method	    
Topic Information	Topic_vision: Wedding; Topic-language: Wedding day;
Ground Truth	The wedding day was a success. The bride looked beautiful. They were both so very happy together. The flower colors looked just perfect with her dress. They make a cute couple.
PR-VIST	[female] and [male] took a family photo together. it was [male] 's birthday party. the groom and his wedding were cut off. their mom was so proud of them. her mom got married today, but [female] was very excited to be there. [female] is now getting married today .
KE-VIST	the bride and groom were getting ready for their wedding. they looked so happy to be there. it was a beautiful night. everyone danced. then came out after .
TARN-VIST	the bride and groom are married. The bride is taking a picture with her friends. she was happy and her husband was so happy. after the wedding, we had a great time with her wedding. they were very excited to get a picture of the wedding.

Conclusion and Future Work

Conclusion

- Mine the story's topic information from both visual and linguistic perspectives
- Employ the reinforcement learning to refine the generation process
- Topic information is beneficial to improve the quality of generated stories

Future Work

- Explore grammar and discourse structure in the visual storytelling task
- Analyze linguistic style to improve the quality and diversity of generated stories

Thank you!