Video and Language Description Targeted Person Search

Speaker: Shuang Li
Supervisor: Prof. Xiaogang Wang
Prof. Hongsheng Li

Person Search

Person Re-identification

Person Search

- **Image** based person re-identification/search

- **Video** based person search

- **Natural language** based person search

The woman is wearing a long, bright orange gown with a white belt at her waist.

Video-based Person Search

Diversity Regularized Spatiotemporal Attention for Video-based Person Search

Challenges

Pose variation

Viewpoints

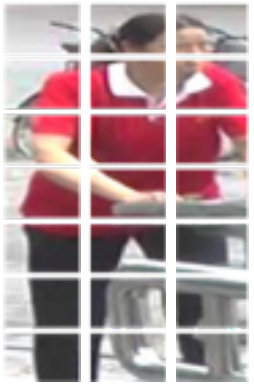
Occlusion

Low illumination

Low resolution

Background clutter

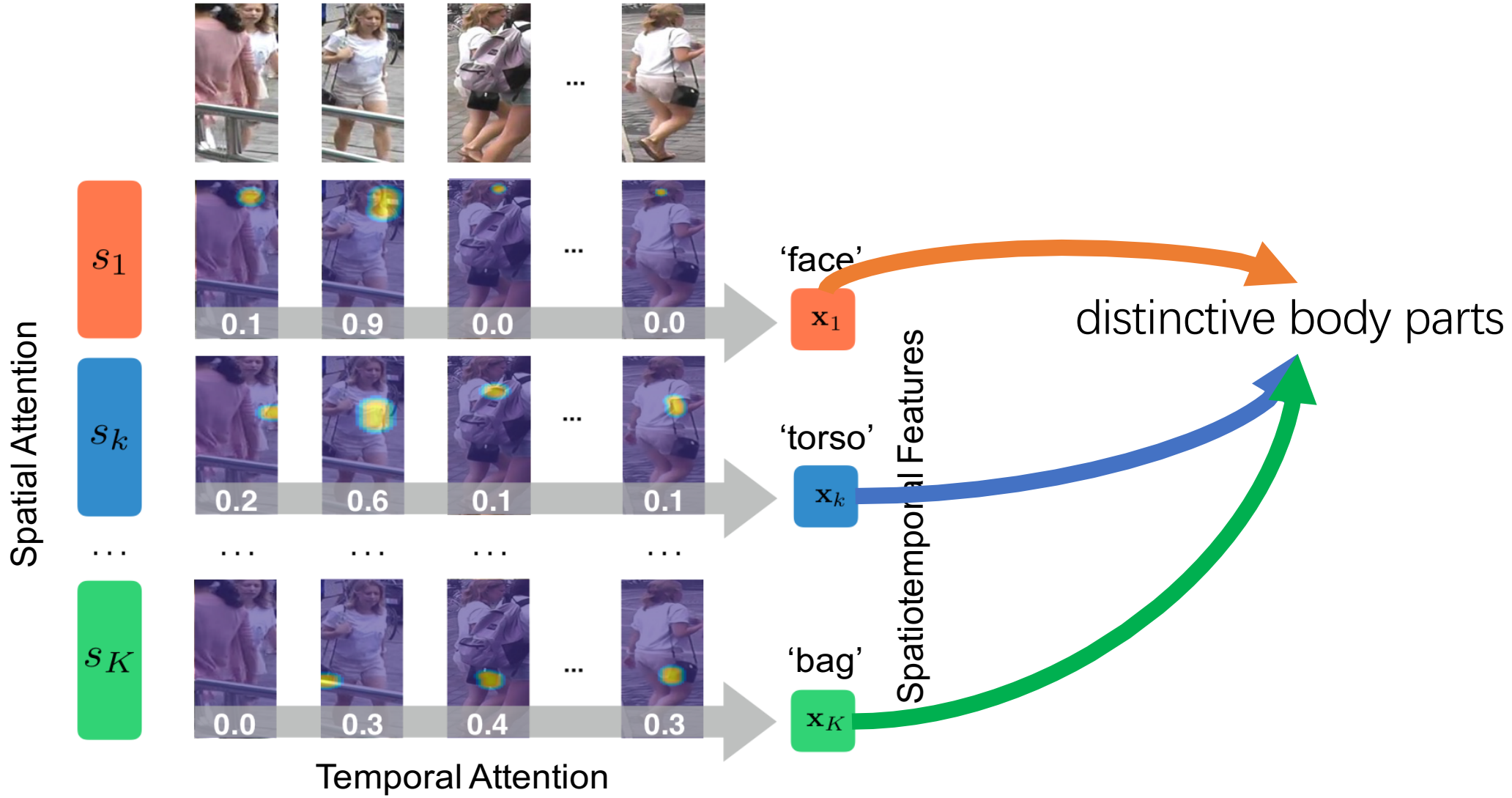
Challenges

Spatial	Input	Limitations
Whole image		Features are corrupted by occlusions; Miss fine-grained visual cues;
Predefined compositions (grid or body parts)		Rough grid separation is unreliable; No body part annotations; Hard to predefine accessories;

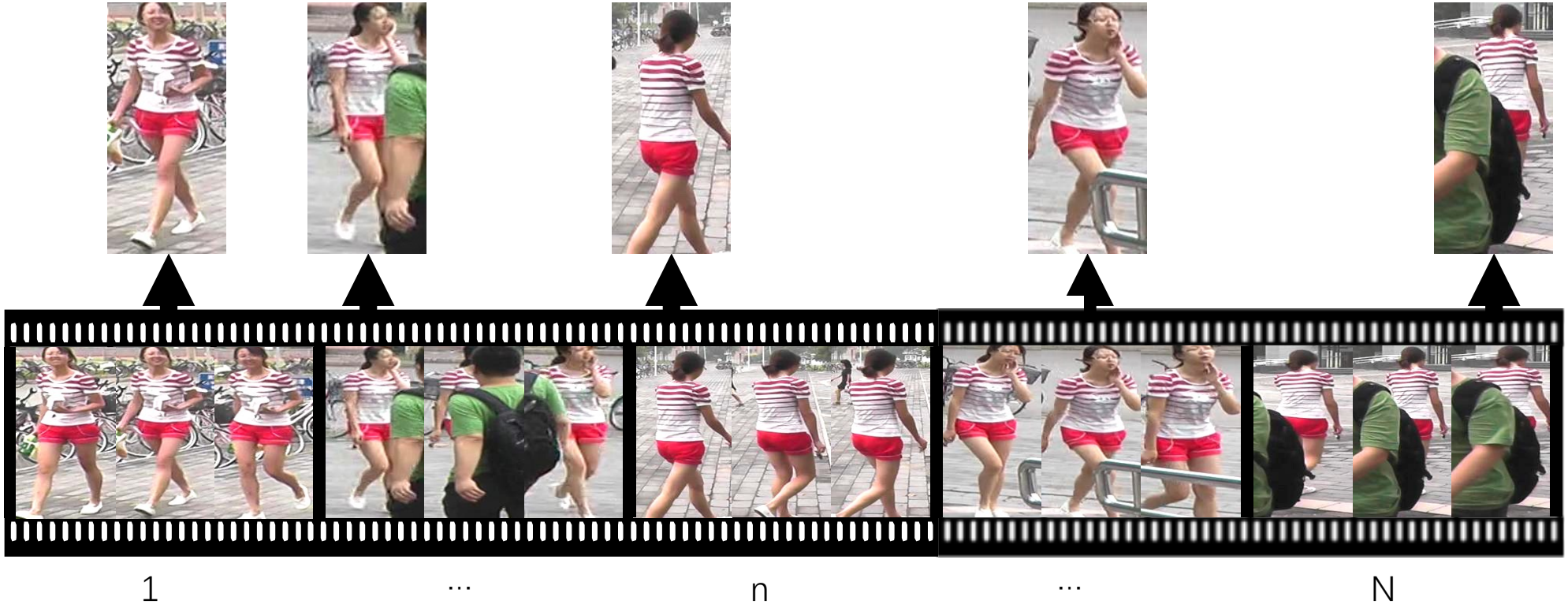
Challenges

Temporal	Limitations
	Person pose change over time; ignore spatial misalignment when comparing frame features; Temporal aggregation (avg/max pooling) generally weaken or over emphasize the contribution of discriminative frames.

Spatiotemporal Attention

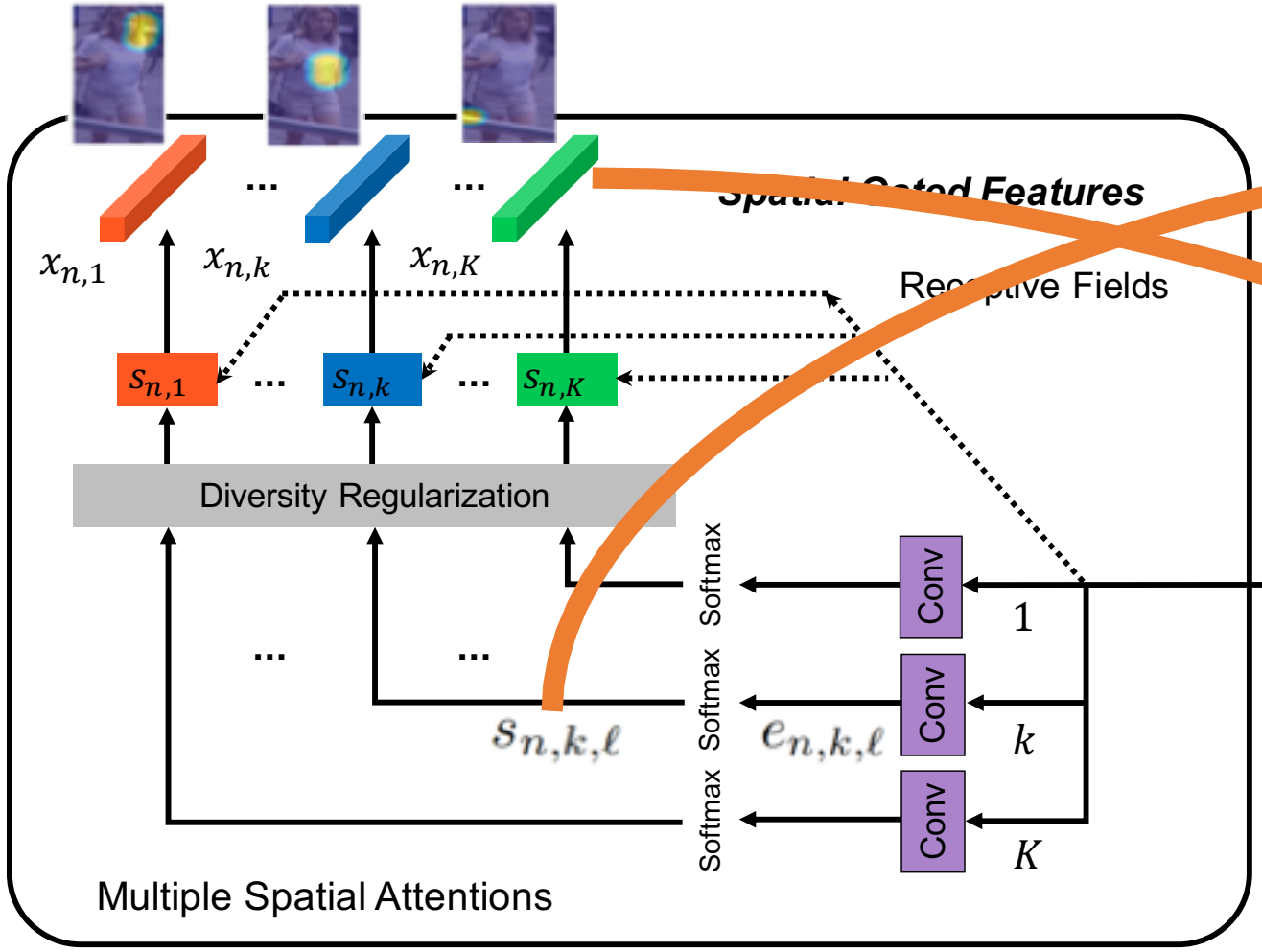


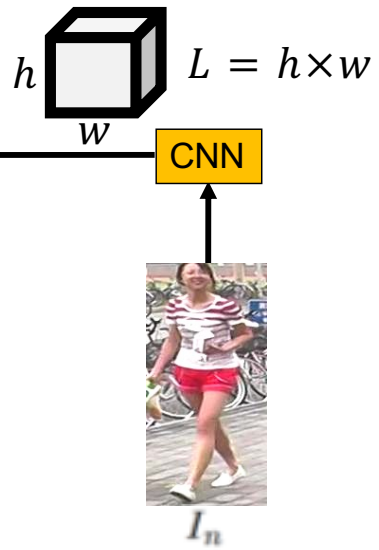
Restricted Random Sampling



- Consecutive video frames are highly corelated
- Significant visual changes over time

Multiple Spatial Attention





Attention value:

$$s_{n,k,l} = softmax(e_{n,k,l})$$

$$\sum_{l=1}^{L} s_{n,k,l} = 1$$

Receptive field:

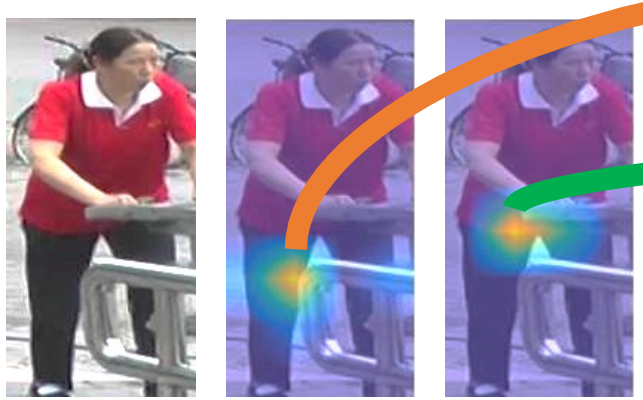
$$s_{n,k} = [s_{n,k,1}, \ldots, s_{n,k,L}]$$

Spatial gated feature:

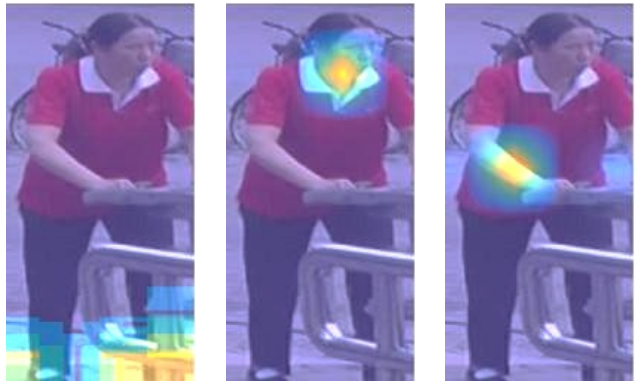
$$x_{n,k} = \sum_{l=1}^{L} s_{n,k,l} f_{n,l}$$

$$k = [1, \ldots, K]$$

Diversity Regularization



knee hip



foot neck waist

Attention vectors (receptive fields): $s_{n,i}, s_{n,j}$

Hellinger distance:

$$H(s_{n,i}, s_{n,j}) = \frac{1}{\sqrt{2}} \left\| \sqrt{s_{n,i}} - \sqrt{s_{n,j}} \right\|_2 = \frac{1}{\sqrt{2}} \sqrt{\sum_{l=1}^{L} (\sqrt{s_{n,i,l}} - \sqrt{s_{n,j,l}})^2}$$

Since $\sum_{l=1}^{L} s_{n,i,l} = 1, \sum_{l=1}^{L} s_{n,j,l} = 1,$

$$S_n = [s_{n,1}, \dots, s_{n,K}], R_n = \sqrt{S_n}$$

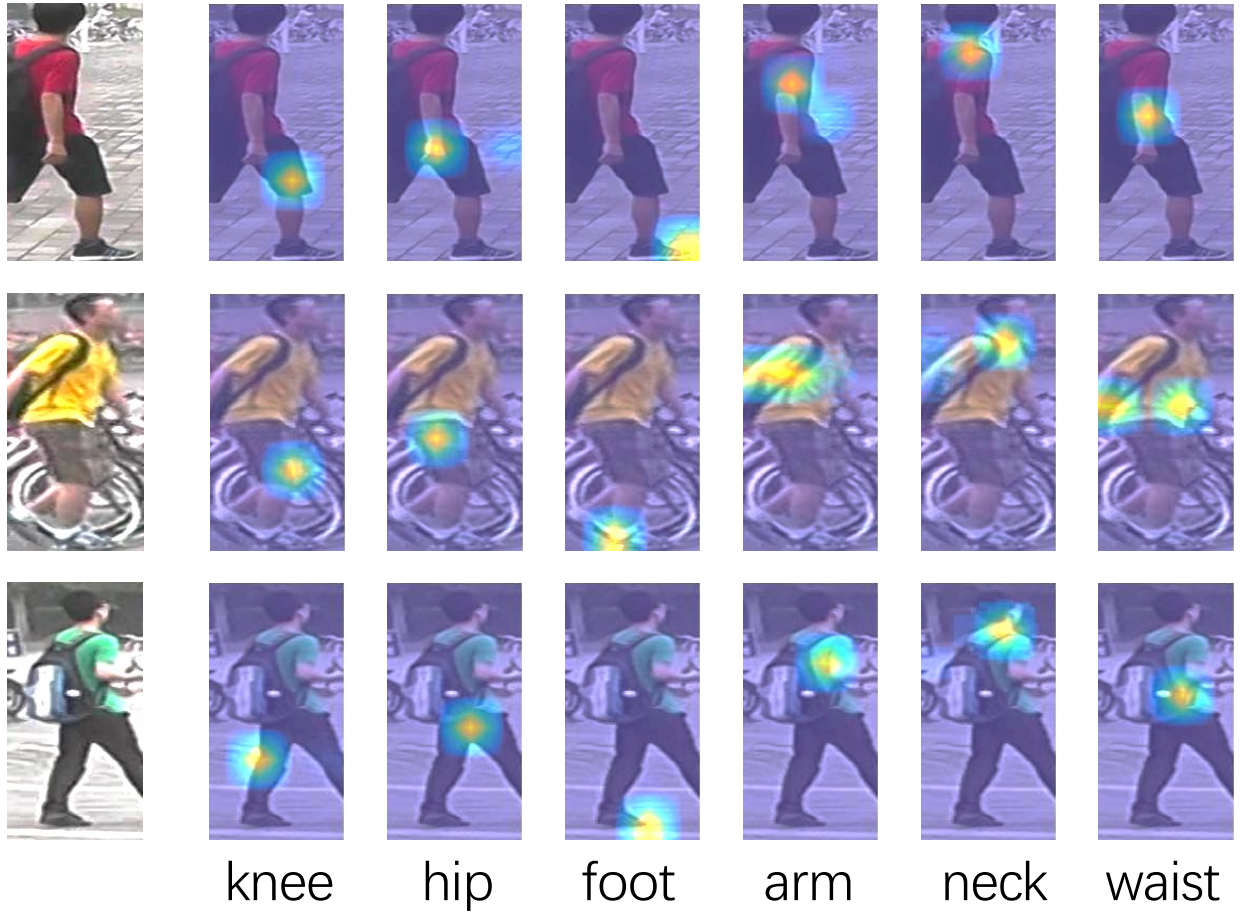
$$H^2(s_{n,i}, s_{n,j}) = 1 - \sum_{l=1}^{L} (\sqrt{s_{n,i,l} s_{n,j,l}})$$

$$Q = \|(R_n R_n^T - I)\|_F^2$$

Maximize distance between any two attention vectors $s_{n,i}$ and $s_{n,j}$

Minimize $1 - H^2(s_{n,i}, s_{n,j}) = \sum_{l=1}^{L} (\sqrt{s_{n,i,l} s_{n,j,l}})$

Learned Spatial Attention Models

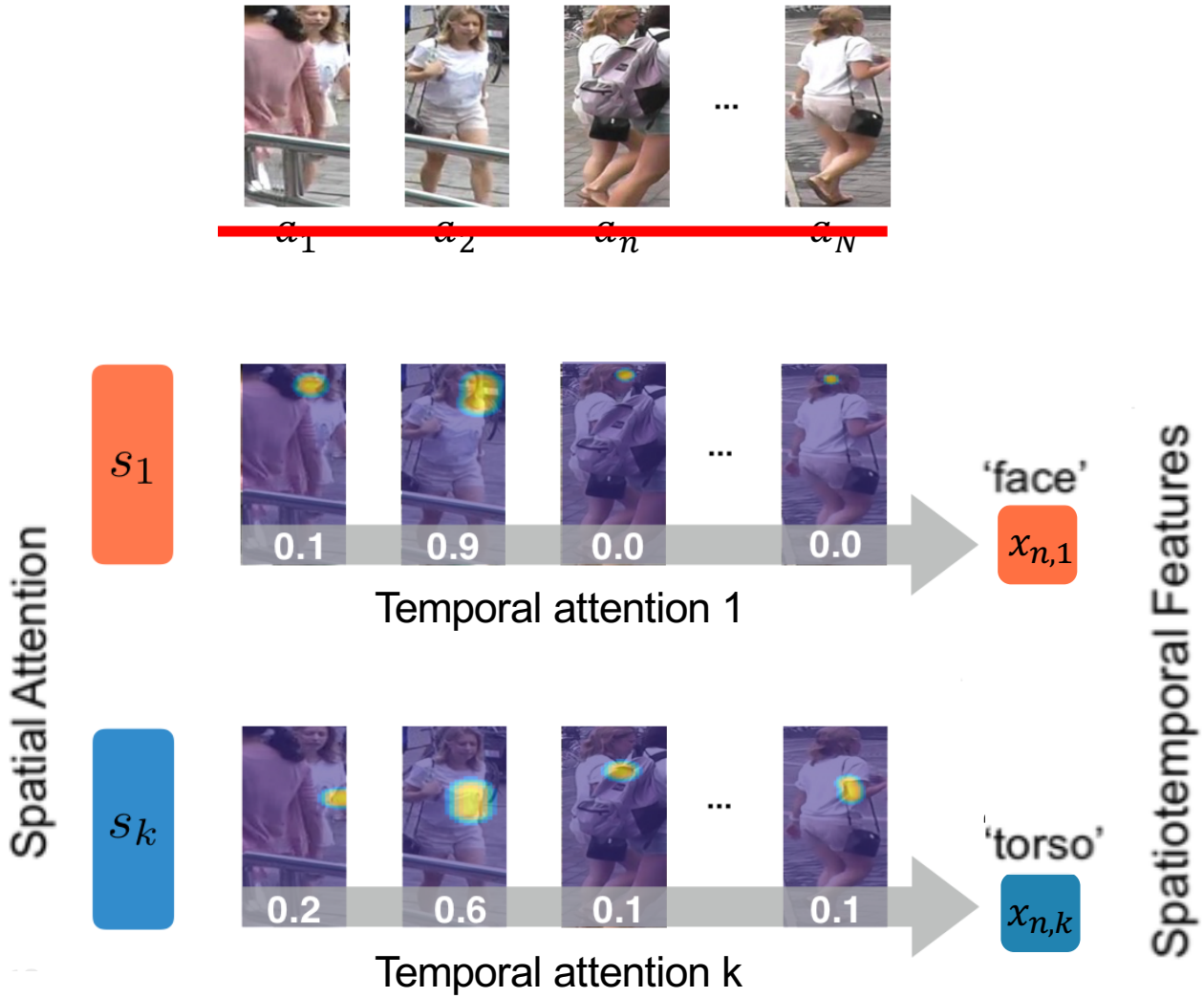


Examples of 6 spatial attention models after applying the diversity regularization.

Challenges

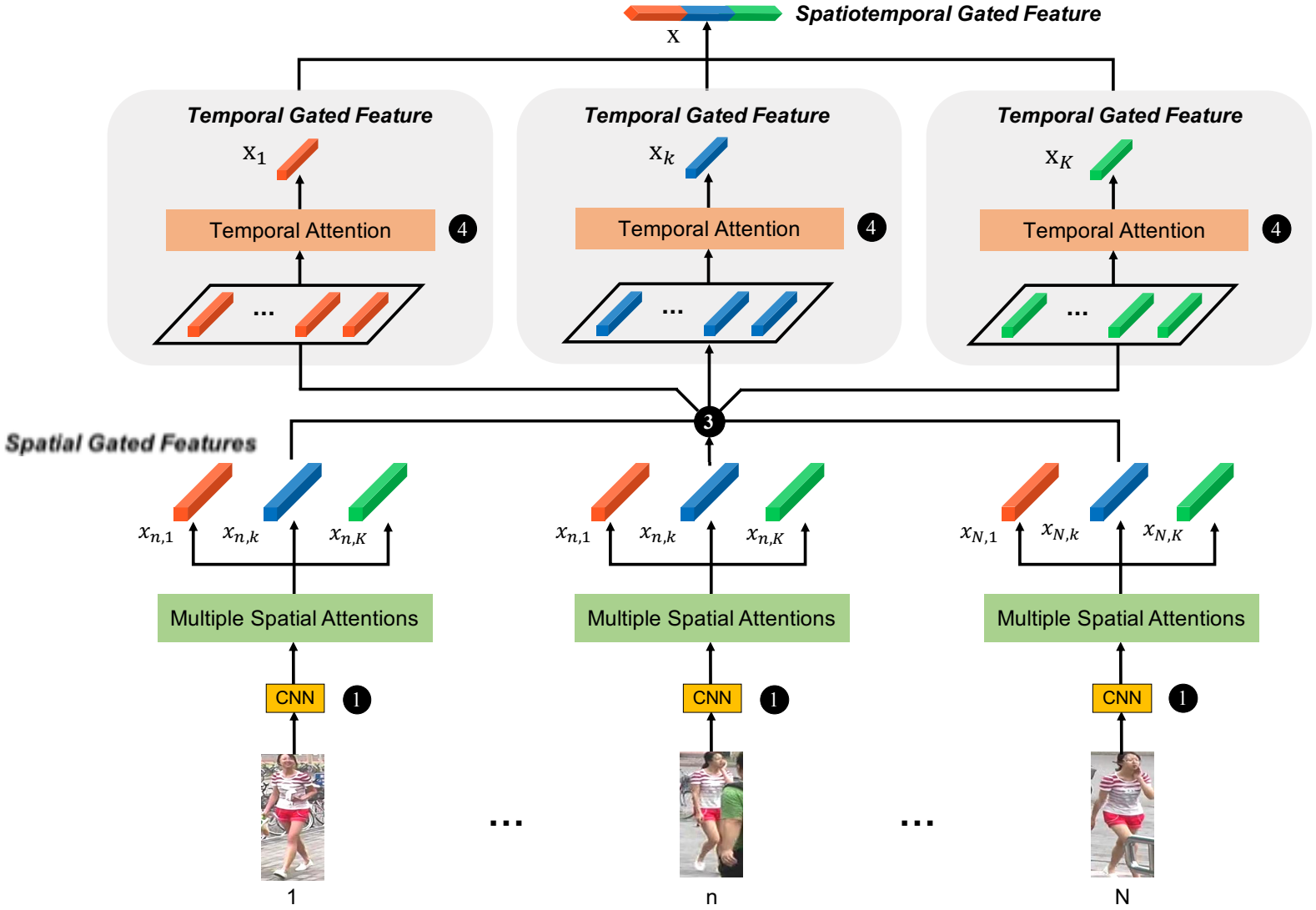
	input	Limitations
Spatial	Whole image	Features are corrupted by occlusions; Miss fine-grained visual cues;
	Predefined compositions	Rough grid separation is unreliable; No body part annotations; Hard to define components;
Temporal	Video sequence	Person pose change over time; ignore spatial misalignment when comparing frame features; Temporal aggregation generally weaken or over emphasize the contribution of discriminative frames.

Temporal Attention



- All parts of an object are seldom visible in every video frame;
- Occluded frames contain valuable partial information, e.g. face;
- A single temporal attention model on the whole frame features could miss fine-grained visual information;
- Multiple temporal attention weights.

Temporal Attention



$$e_{n,k} = Conv(x_{n,k})$$

Temporal attention value:

$$t_{n,k} = softmax(e_{n,k})$$

Spatiotemporal gated feature:

$$x_k = \sum_{n=1}^{N} t_{n,k} x_{n,k}$$

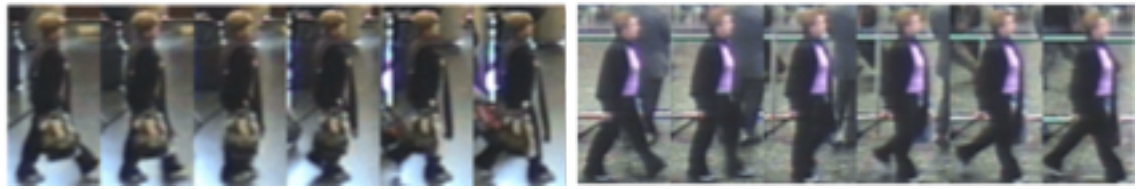
$$x = [x_{1,..,} x_K]$$

Challenges

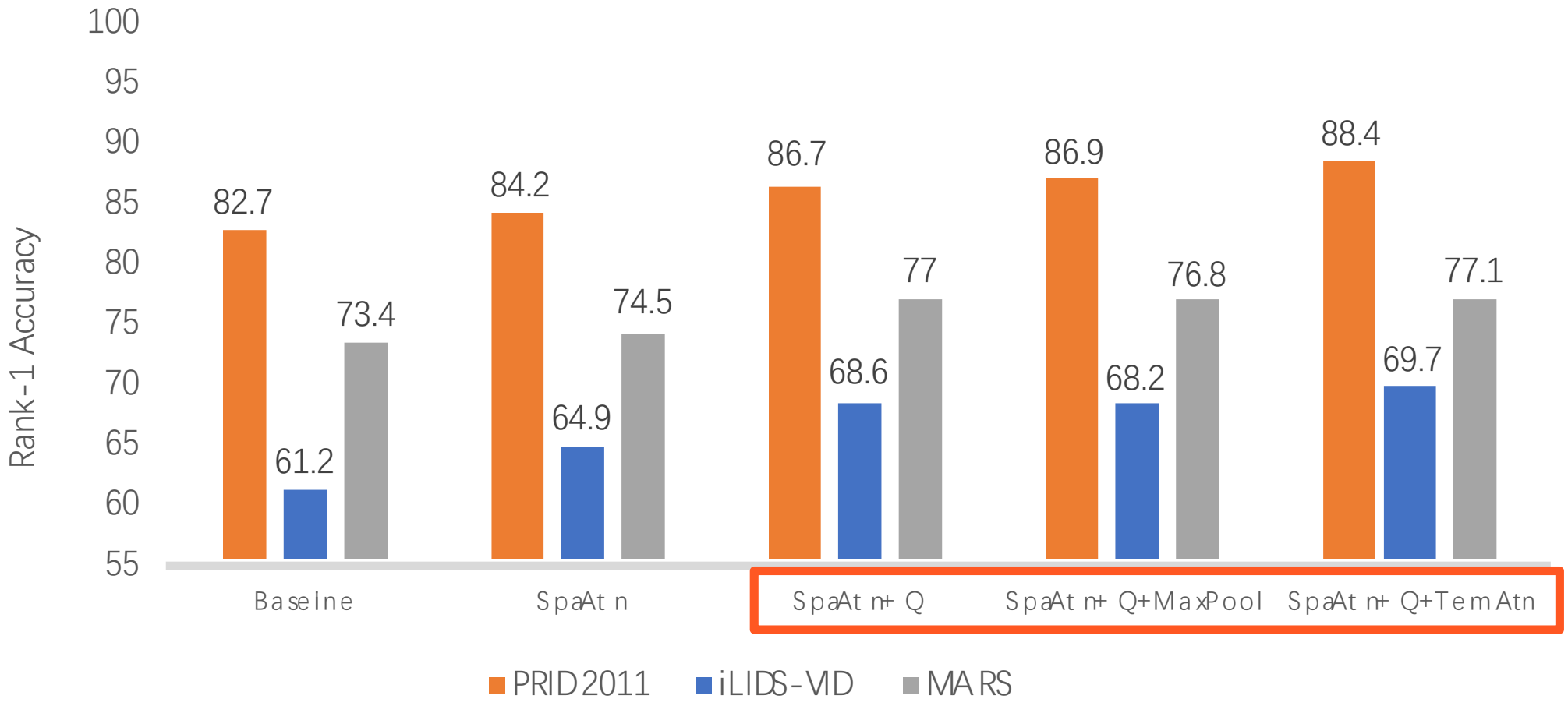
	input	Limitations
Spatial	Whole image	Features are corrupted by occlusions; Miss fine-grained visual cues;
	Predefined compositions	Rough grid separation is unreliable No body part annotations; Hard to define components;
Temporal	Video sequence	Person pose change over time; ignore spatial misalignment when comparing frame features; Temporal aggregation generally weaken or over emphasize the contribution of discriminative frames.

Experiments

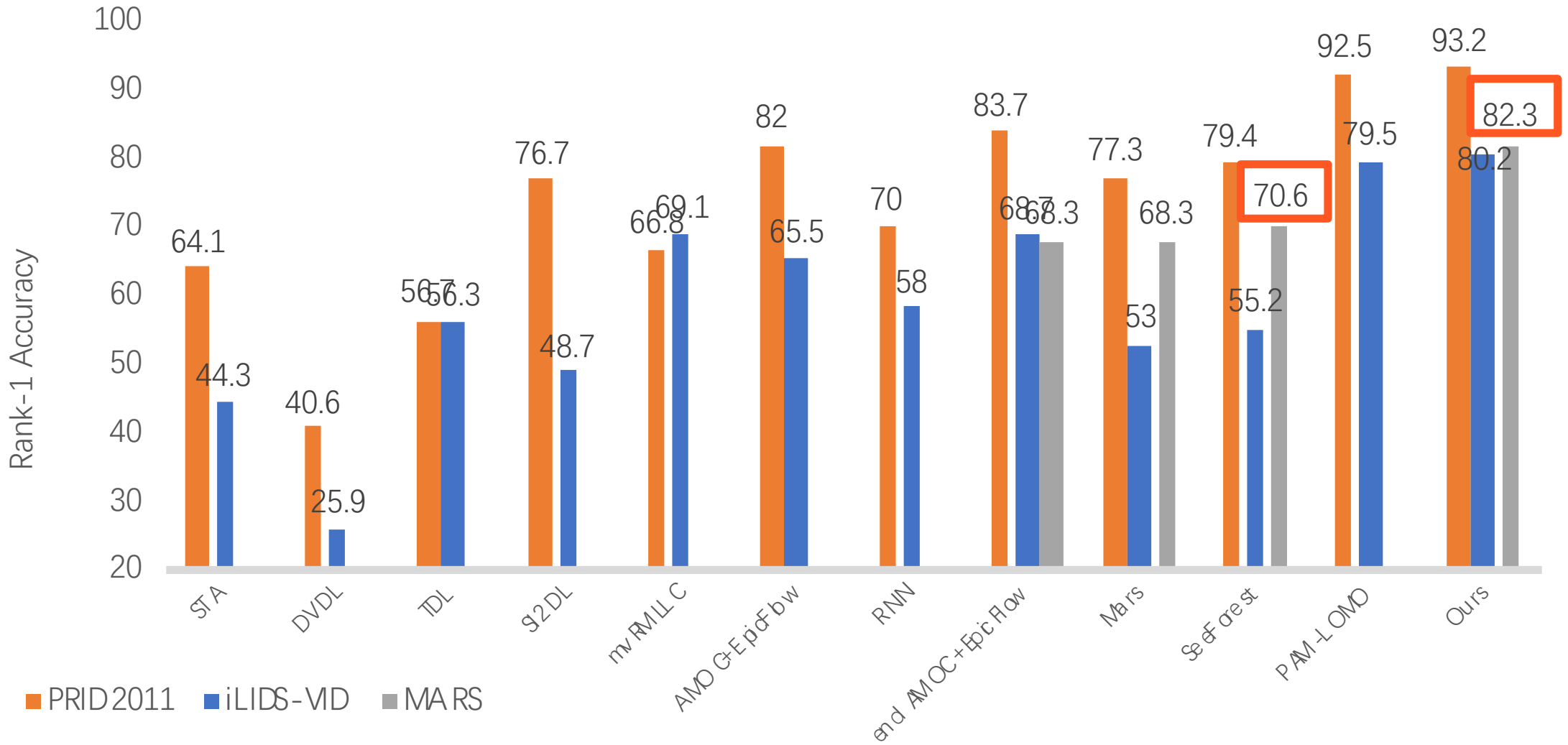
Datasets

Dataset	#id	#vid	#cam	examples
PRID2011	934	1134	2	
iLIDS-VID	300	600	2	
MARS	1261	20,000	6	

Component Analysis



Comparisons with different methods



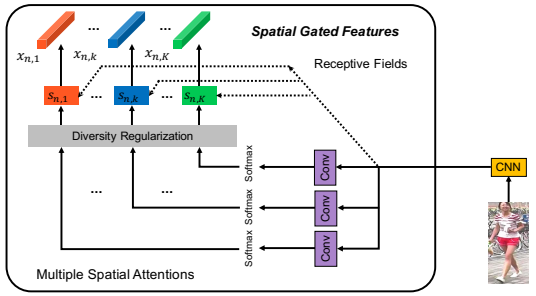
Conclusion

- Multiple spatial attention

  Automatically discover a set of discriminative object parts;

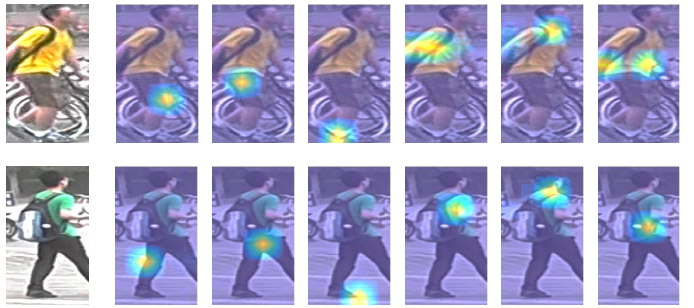
  solve the alignment problem between images;

  avoid features from being corrupted by occluded regions.



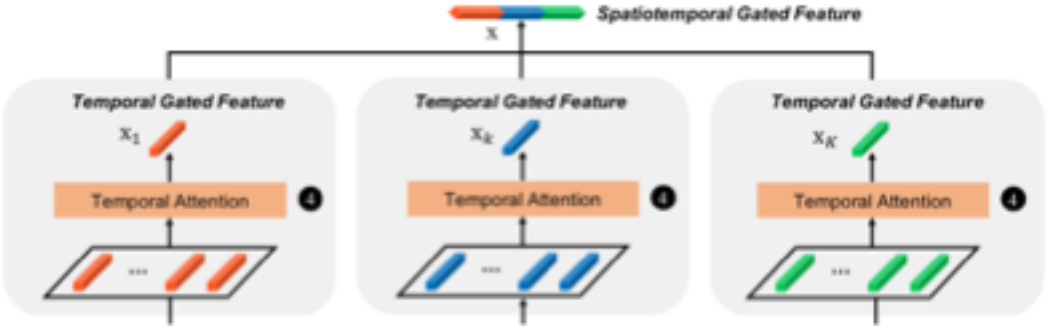
- Hellinger distance diversity regularization

  Ensure multiple spatial attention models do not discover the same body part.



- Multiple temporal attention

  Compute an aggregate representation of the features extracted by each spatial attention model.



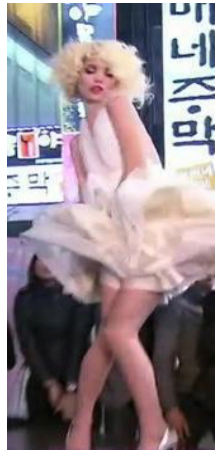
Language-based Person Search

Person Search with
Natural Language Description

Language based Person Search

The woman is wearing a long, bright orange gown with a white belt at her waist. She has her hair pulled back into a bun or ponytail.

CUHK-PEDES Dataset



The woman is dressed up like Marilyn Monroe, with a white dress that is blowing upward in the wind, short curly blonde hair, and high heels.

The woman has long light brown hair, is wearing a black business suit with white low-cut blouse with large, white cuffs, a gold ring, and is talking on a cellphone.

The man is wearing blue scrubs with a white lab coat on top. He is holding paperwork in his hand and has a name badge on the left side of his coat.

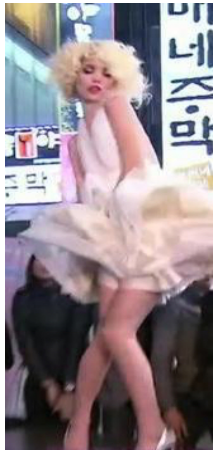
Basic Information

- 40,206 images (from 5 re-id datasets)
- 13,003 persons (3.1 imgs/person)
- 80,412 descriptions (2 descriptions/img)
- Average word length is 23.5
  (Visual Genome 5.18, COCO 10.45)

Workers

- Amazon Mechanical Turk
- 1,993 unique workers
- > 95% approving rate

CUHK-PEDES Dataset

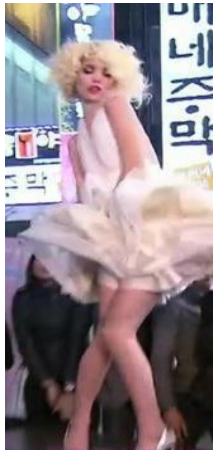


The woman is dressed up like Marilyn Monroe, with a white dress that is blowing upward in the wind, short curly blonde hair, and high heels.

The man is wearing blue scrubs with a white lab coat on top. He is holding paperwork in his hand and has a name badge on the left side of his coat.

- Vocabularies
- Phrases
- Sentence patterns
- Appearances
- Actions
- Poses
- Interactions with other objects.

CUHK-PEDES Dataset

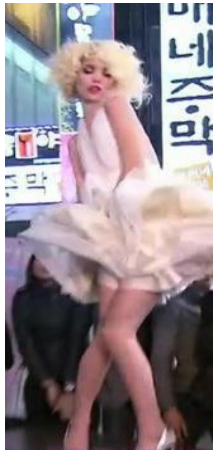


The woman is dressed up like Marilyn Monroe, with a white dress that is blowing upward in the wind, short curly blonde hair, and high heels.

The man is wearing blue scrubs with a white lab coat on top. He is holding paperwork in his hand and has a name badge on the left side of his coat.

- Vocabularies
- Phrases
- Sentence patterns
- Appearances
- Actions
- Poses
- Interactions with other objects.

CUHK-PEDES Dataset

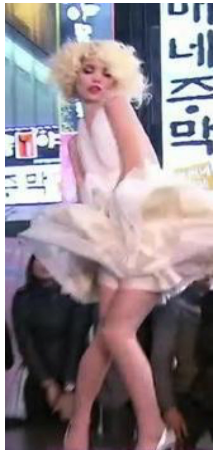


The woman is dressed up like Marilyn Monroe, with a white dress that is blowing upward in the wind, short curly blonde hair, and high heels.

The man is wearing blue scrubs with a white lab coat on top. He is holding paperwork in his hand and has a name badge on the left side of his coat.

- Vocabularies
- Phrases
- Sentence patterns
- Appearances
- Actions
- Poses
- Interactions with other objects.

CUHK-PEDES Dataset

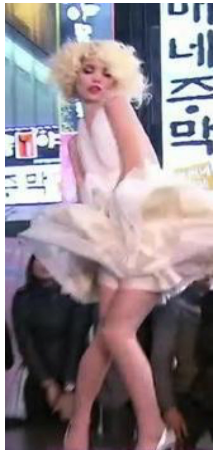


The woman is dressed up like Marilyn Monroe, with a white dress that is blowing upward in the wind, short curly blonde hair, and high heels.

The man is wearing blue scrubs with a white lab coat on top. He is holding paperwork in his hand and has a name badge on the left side of his coat.

- Vocabularies
- Phrases
- Sentence patterns
- Appearances
- Actions
- Poses
- Interactions with other objects.

CUHK-PEDES VS COCO



The woman is dressed up like Marilyn Monroe, with a white dress that is blowing upward in the wind, short curly blonde hair, and high heels.

The man is wearing blue scrubs with a white lab coat on top. He is holding paperwork in his hand and has a name badge on the left side of his coat.

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building

COCO
Common Objects in Context

User Study

Carefully read given sentences or attributes and select 5 corresponding images in order

Please annotate at most 50 hits, more hits will be rejected

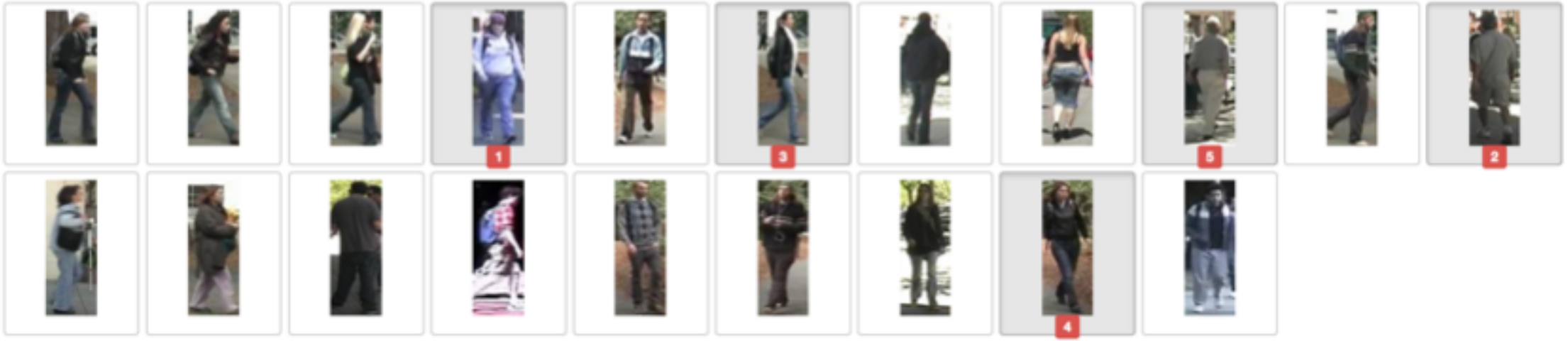
- Read sentences or attributes.
- Select 5 images in increasing order of correspondence

Please read the example:

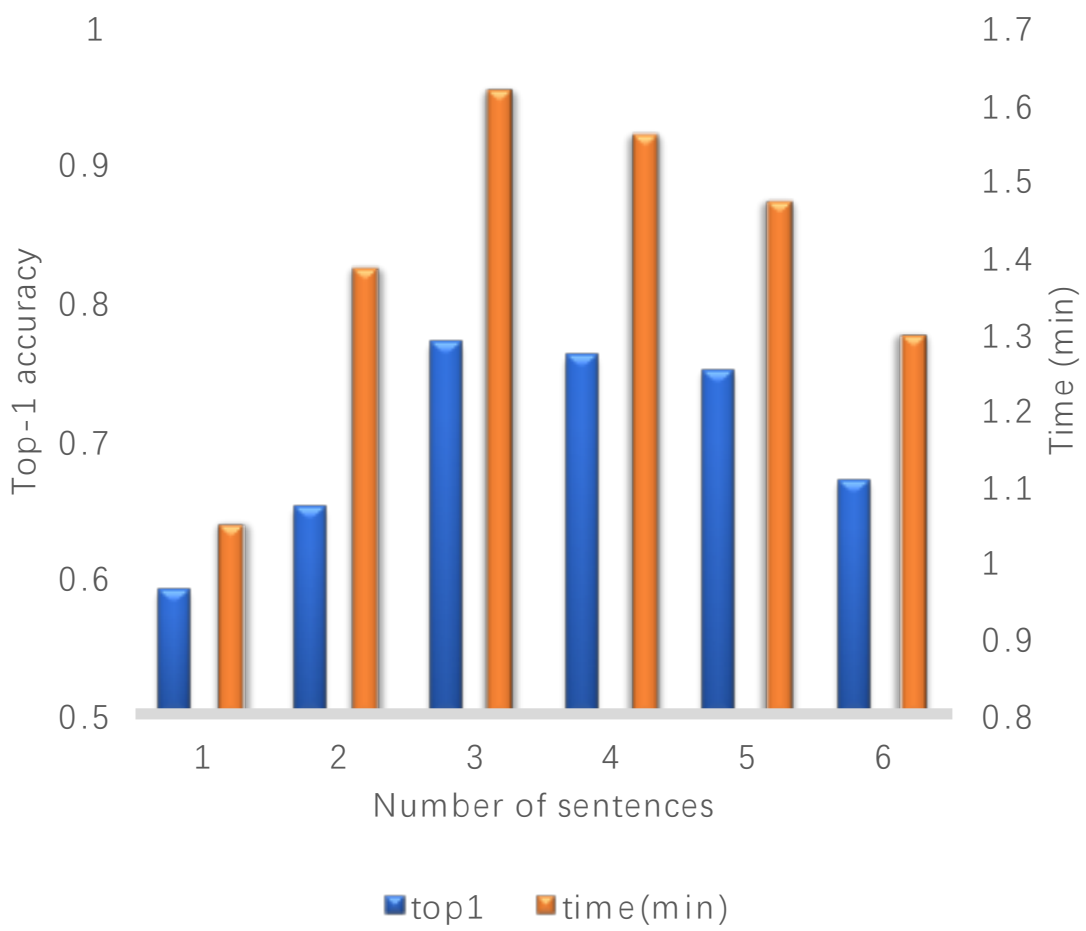
Sentences:

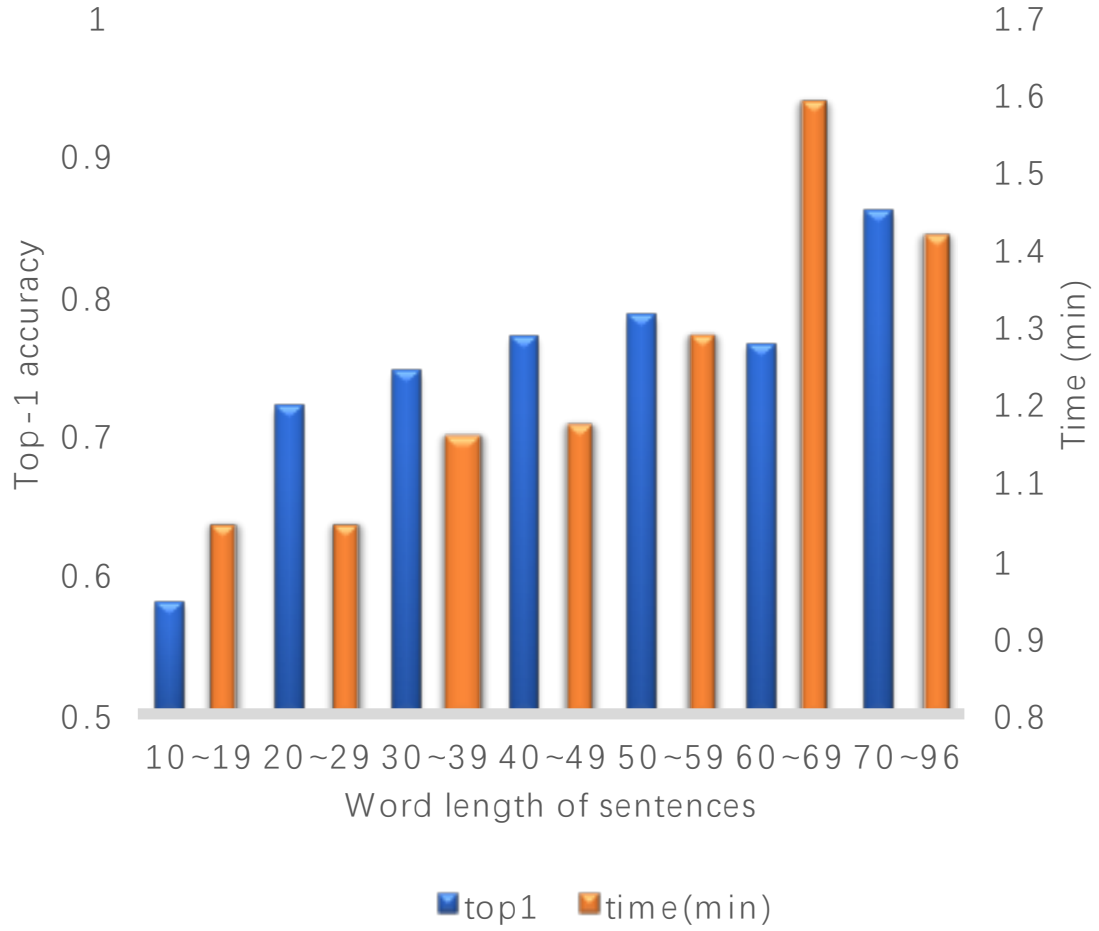
- The woman is wearing a white hoodie with blue pants and grey shoes.She is also carrying a black backpack and wears a burgundy type of hat
- A woman is wearing a purple hat, light colored sweatshirt, light blue jeans and sneakers.

Select 5 images 5/5 **:**



Sentence Number and Length





Word Type

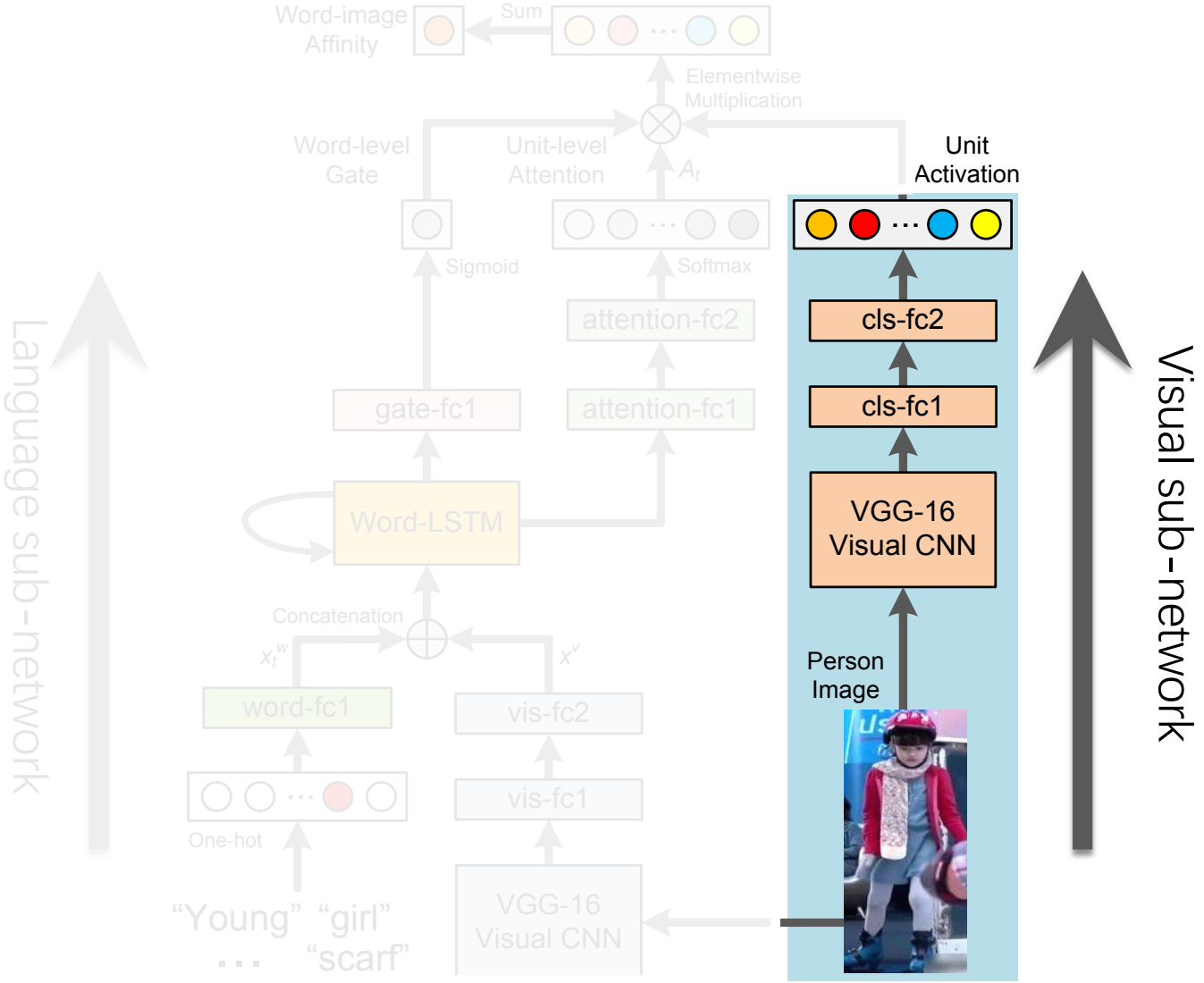
“the girl has brown hair”

“the **** has brown ****”

	orig. sent	w/o nouns	w/o adjectives	w/o verbs
Top-1	0.34	0.23	0.26	0.33
Time (s)	1.14	1.01	0.94	1.13

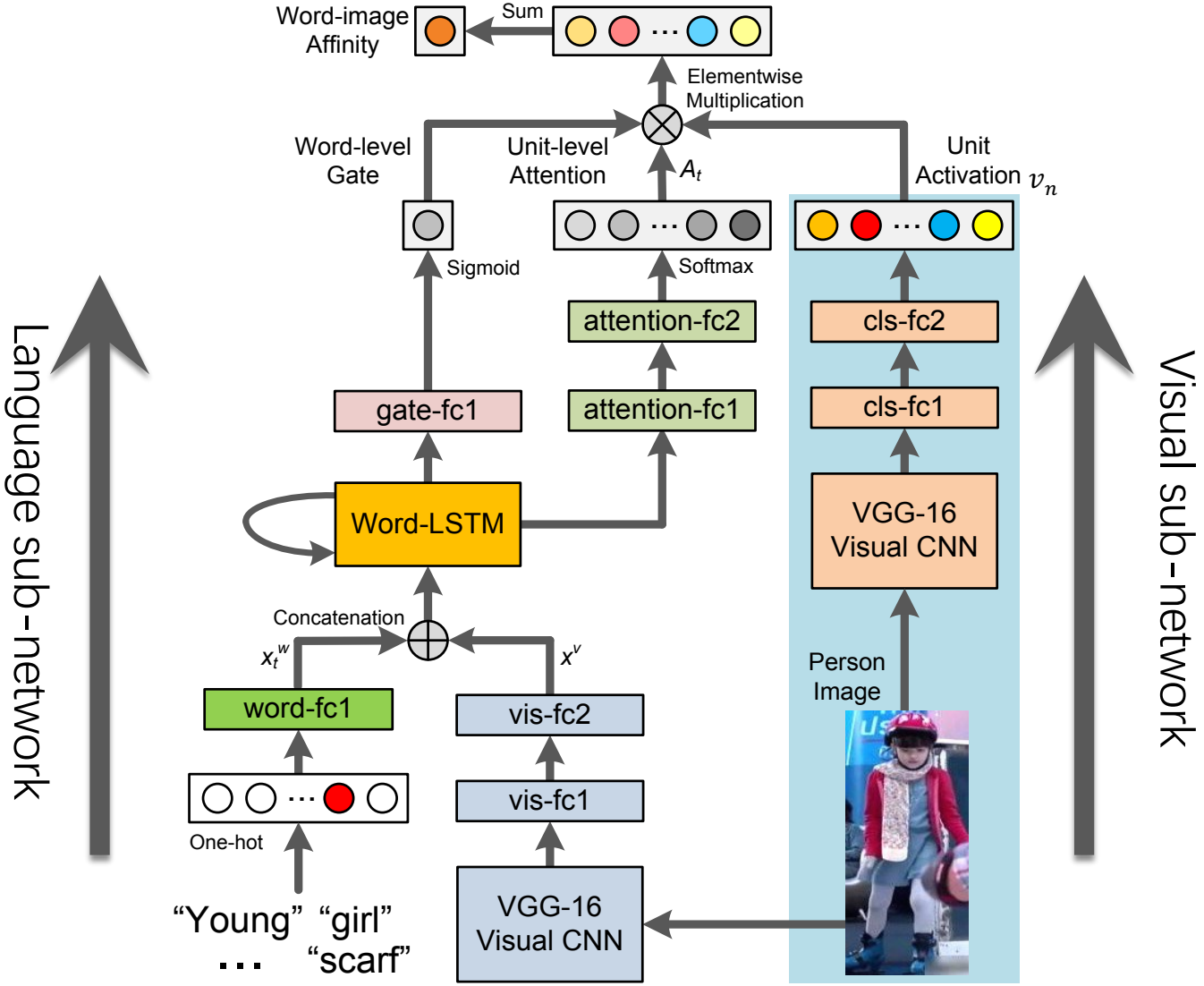
Deep Recurrent Neural Network with Gated Neural Attention

Visual Units



[1] Interpreting Deep Visual Representations via Network Dissection, D. Bau, B. Zhou, A. Oliva, and A. Torralba.

Attention over Visual Units



Word-image affinity

$$a_t = \sum_{n=1}^{N} A_t(n) v_n$$

Sentence-image affinity

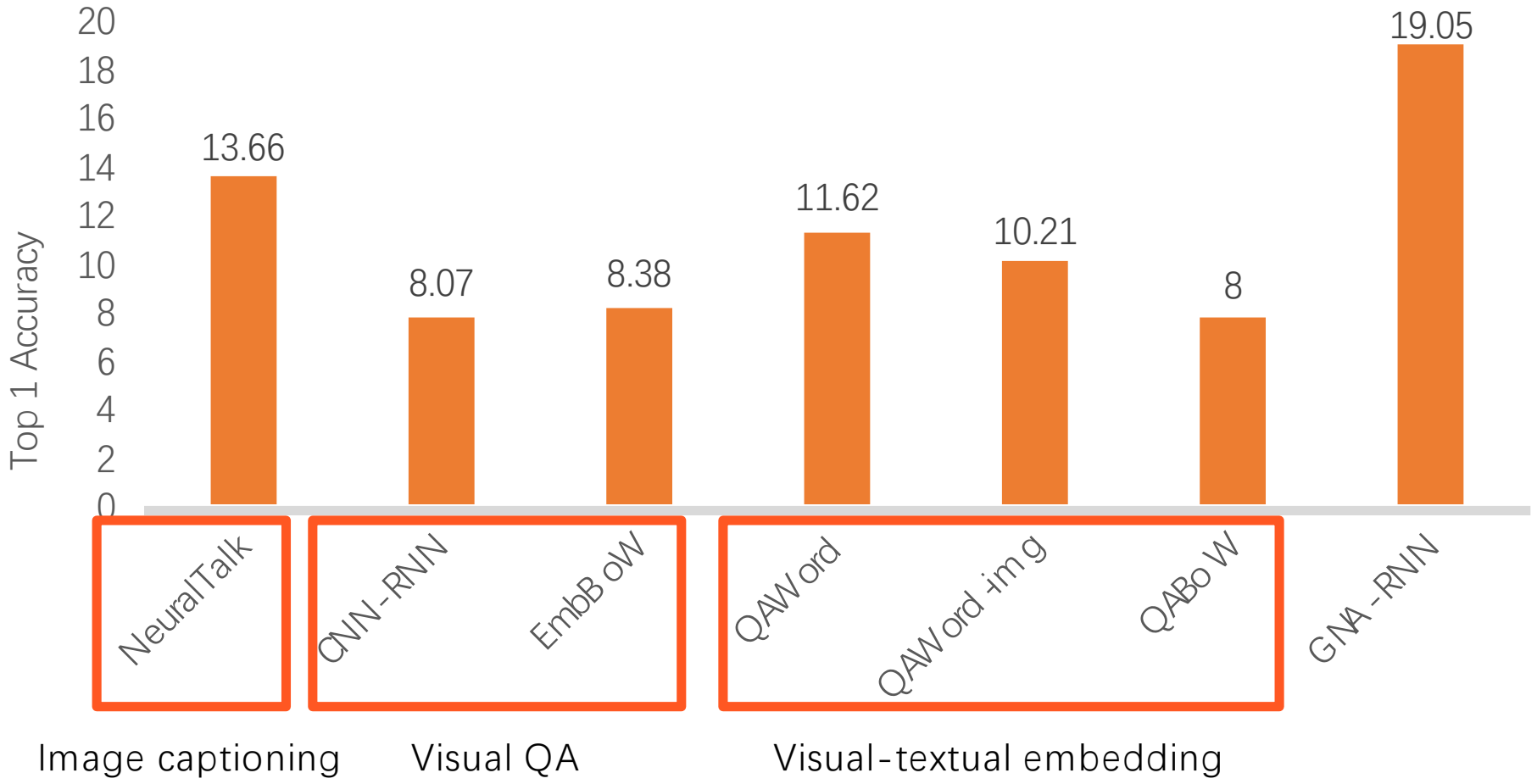
$$a = \sum_{t=1}^{T} a_t$$

Gated sentence-image affinity

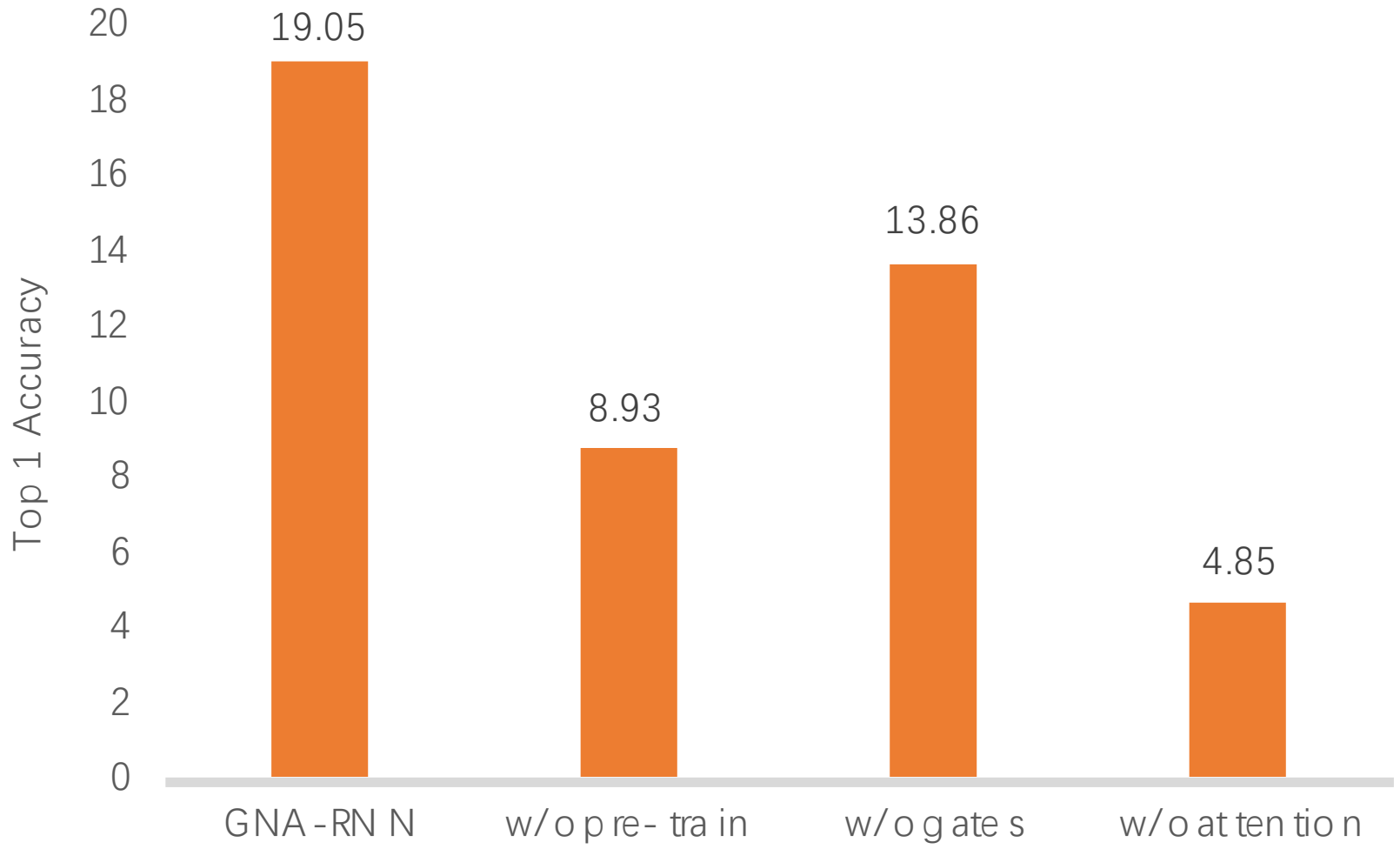
$$a_t = g_t \sum_{n=1}^{N} A_t(n) v_n$$

Experiments

Comparisons with different methods

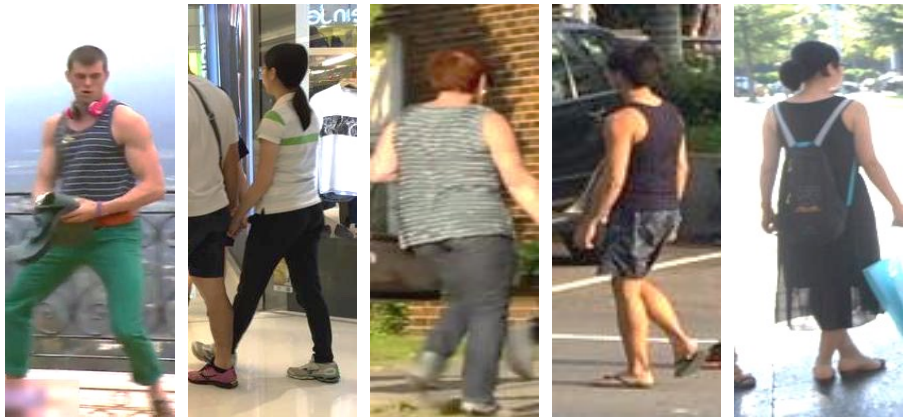


Component Analysis



Visual Unit Visualization

backpack (unit 5)



sleeveless (unit 24)

pink (unit 149)

yellow (unit 205)

Qualitative Evaluation

The woman is wearing a white wedding dress with brown hair pulled back into a long white veil. The dress is cinched with a white ribbon belt.

A woman is wearing a bright red shirt, a pair of black pants and a pair of black shoes.

A man has short brown hair and glasses. He wears a grey suit with a white collared shirt and black tie. He carries a white binder.

The woman is wearing a white top and khaki skirt. She carries a red hand bag.

Conclusion

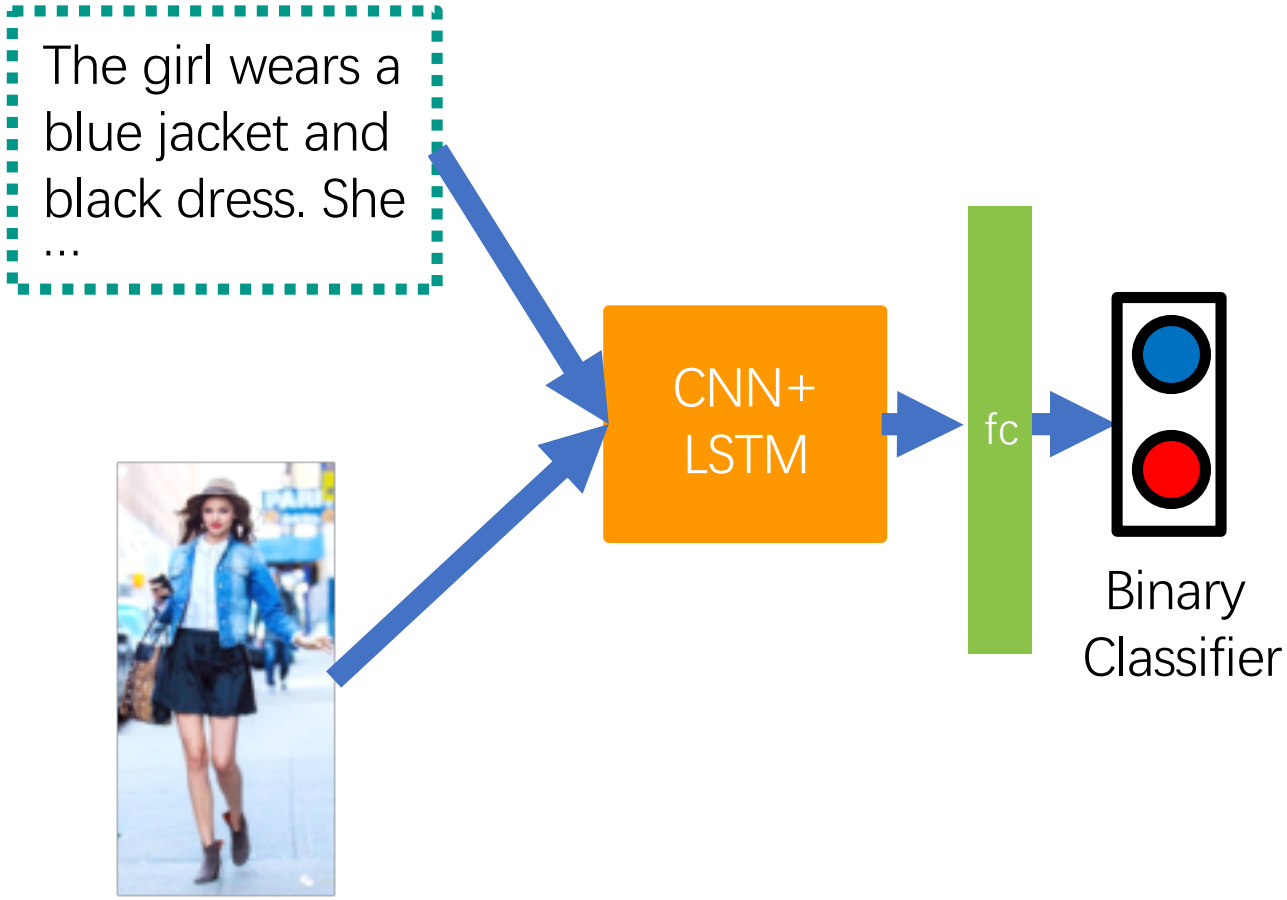
- Propose the problem of person search with natural language description.
- Collect a large-scale person description dataset with rich language annotations.
- Propose a Recurrent Neural Network with Gated Neural Attention for person search.
- Investigate a wide range of plausible solutions and establish baselines on the person search benchmark.

Textual-Visual Matching

From Natural Language based Person Search to Textual-Visual Matching

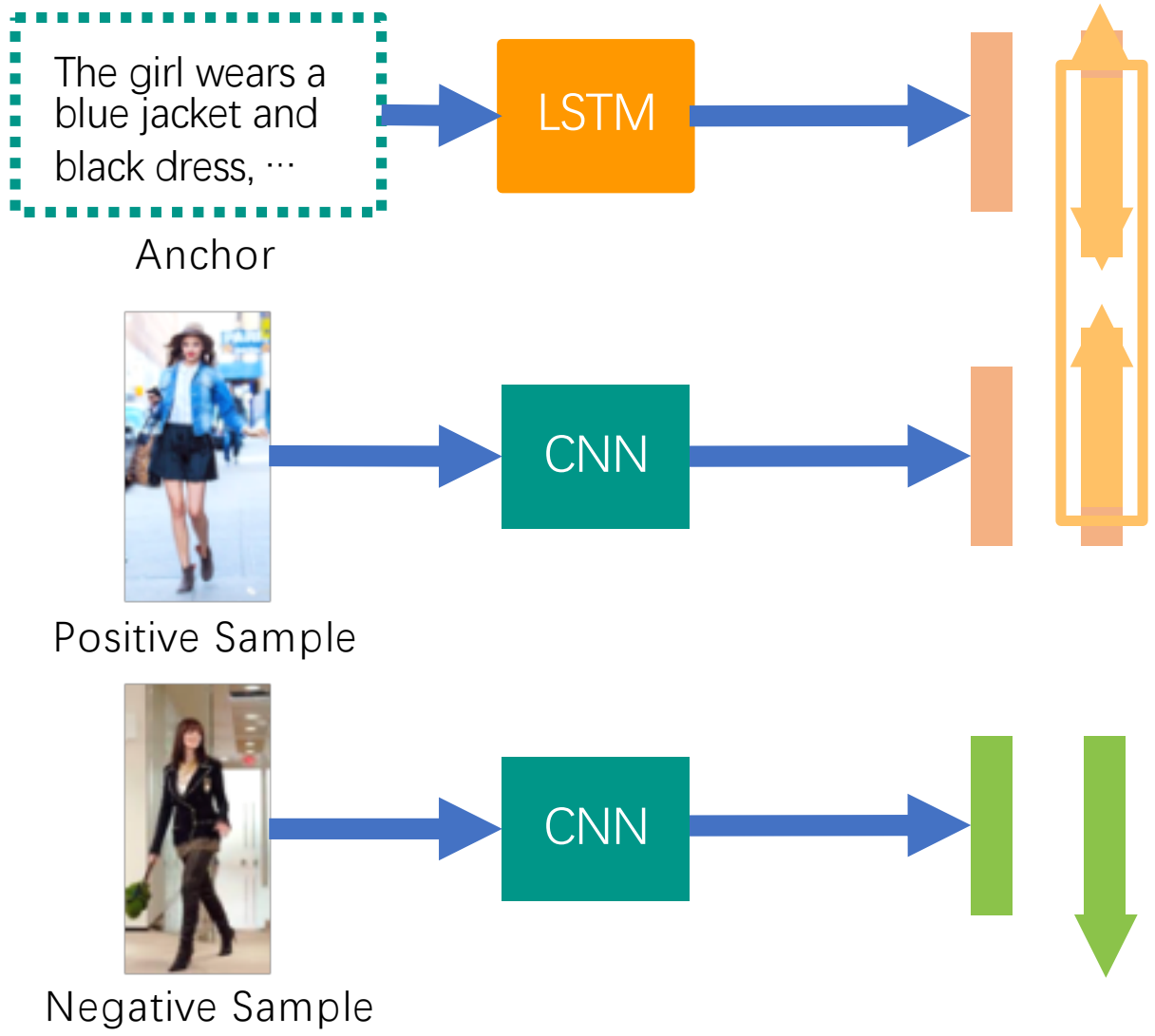
Possible Solutions

- Pair-wise Loss
- Triplet Loss
- Classification Loss



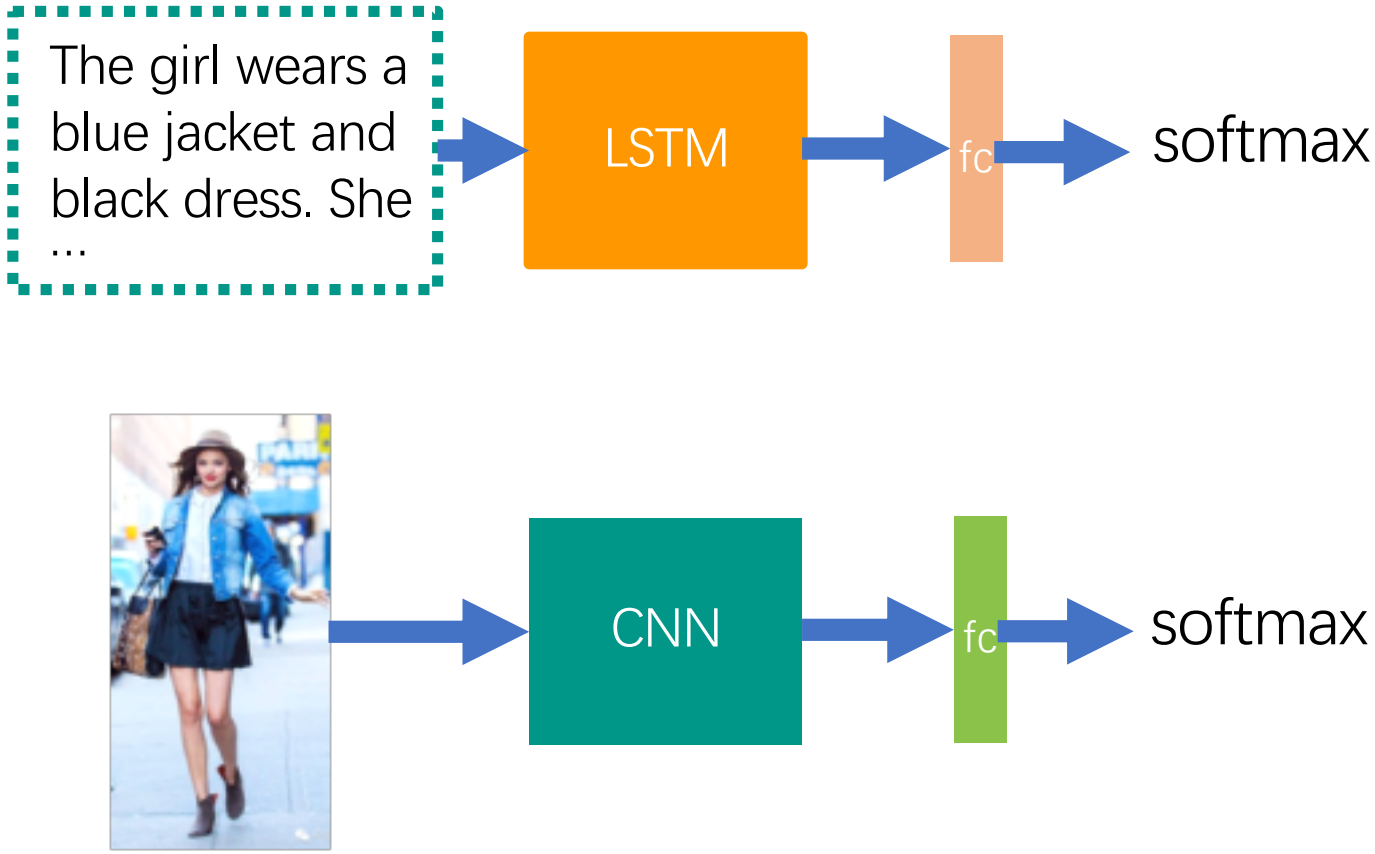
Possible Solutions

- Pair-wise Loss
- Triplet Loss
- Classification Loss



Possible Solutions

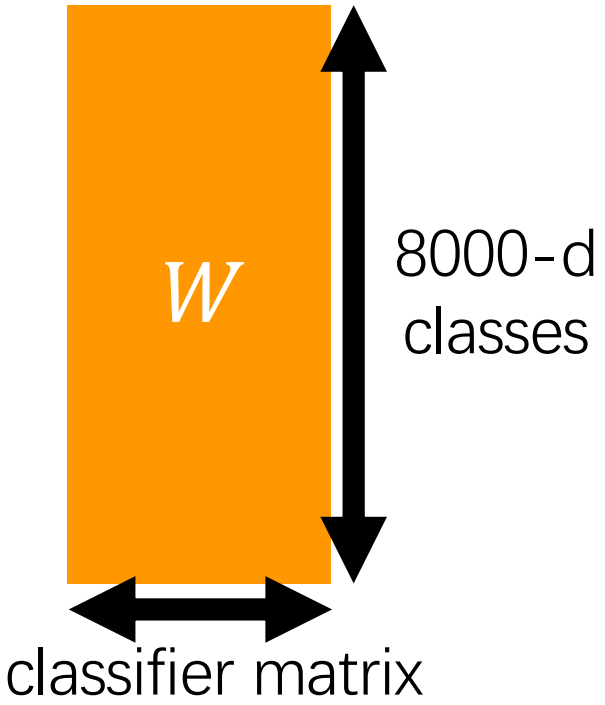
- Pair-wise Loss
- Triplet Loss
- Classification Loss



Limitations of Possible Solutions

Solutions	Phase	Limitations
Pair-wise Loss	Test	N^2 visual-textual pairs
Triplet Loss	Train	Difficulty in selecting hard negative samples
Classification Loss	Train	Few positives for classifier matrix; different feature embedding space

Problems with Classification Loss

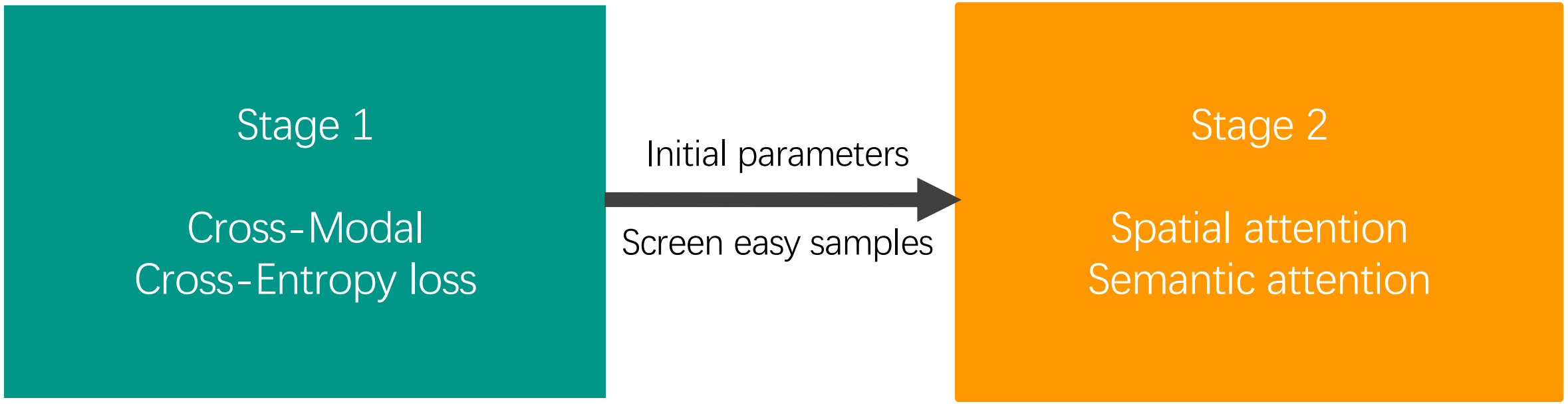


- $p(y = i|x) = \frac{\exp(w_i^T x)}{\sum_j \exp(w_j^T x)}$
- # Positive classes $\leq$ batch size $\ll$ 8000
- No positive samples for most rows
- W cannot be learned effectively

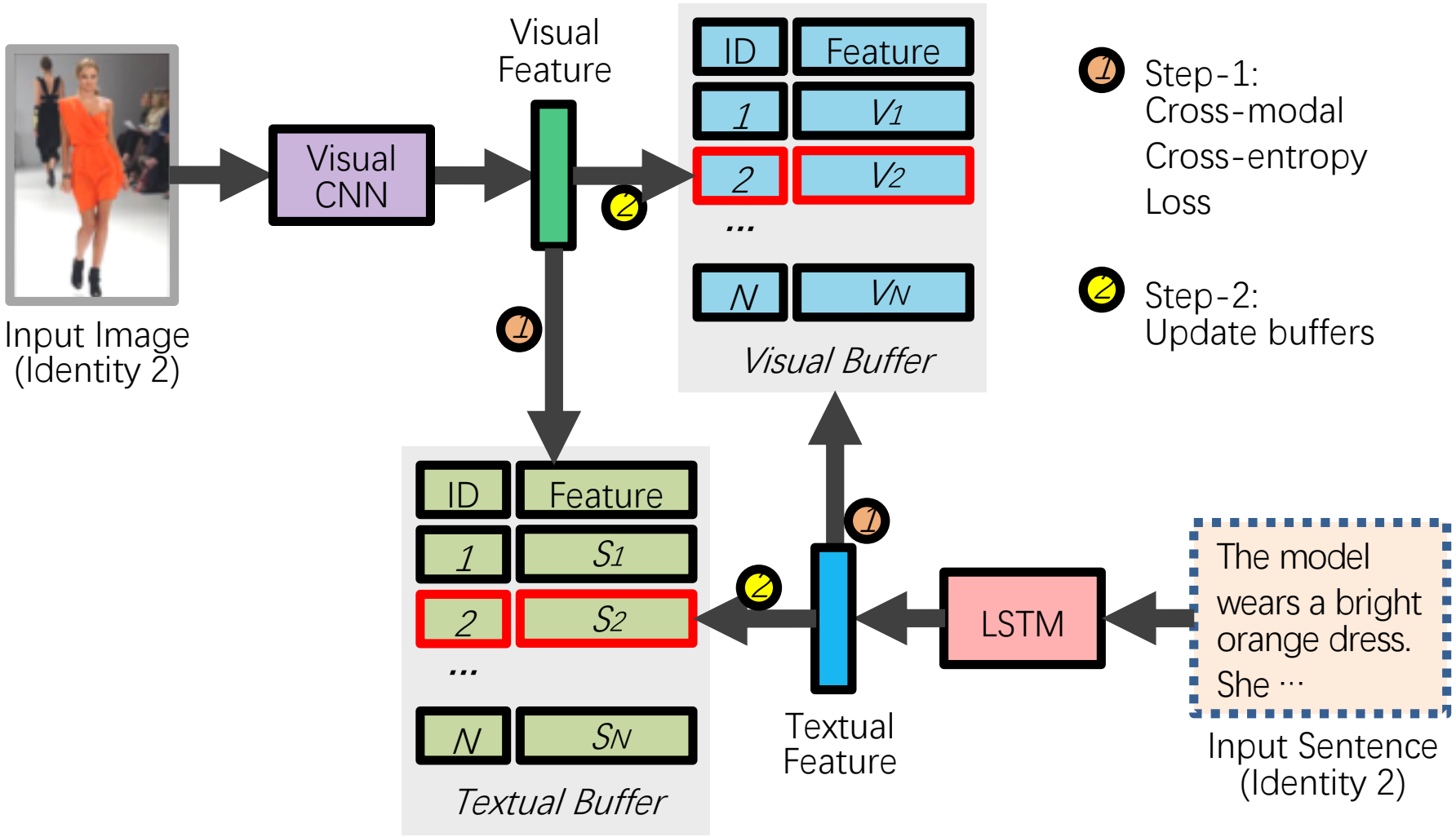
Limitations of Possible Solutions

Solutions	Phase	Limitations
Pair-wise Loss	Test	N^2 visual-textual pairs
Triplet Loss	Train	Difficulty in selecting hard negative samples
Classification Loss	Train	Few positives for classifier matrix; different feature embedding space

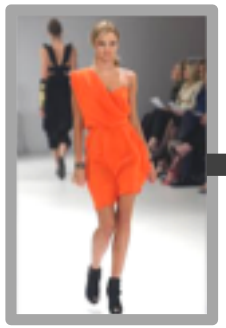
Our framework



Stage-1 CNN-LSTM with CMCE loss



Stage-1 CNN-LSTM with CMCE loss



Input Image
(Identity 2)

Visual CNN

Visual Feature v

ID	Feature
1	V_1
2	V_2
...	
N	V_N

Visual Buffer

ID	Feature
1	S_1
2	S_2
...	
N	S_N

Textual Buffer

Textual Feature s

LSTM

The model wears a bright orange dress. She ···

Input Sentence
(Identity 2)

① Step-1: Cross-modal Cross-entropy Loss

② Step-2: Update buffers

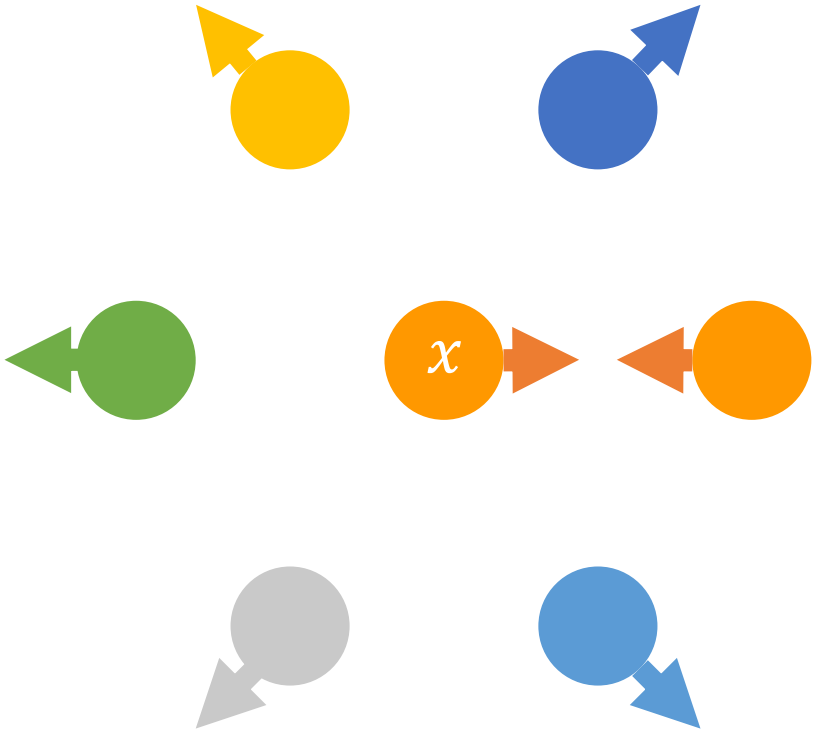
$$p_i^S(v) = \frac{\exp(S_i^T v)}{\sum_{j=1}^{N} \exp(S_j^T v)}$$

$$p_i^V(s) = \frac{\exp(V_i^T s)}{\sum_{j=1}^{N} \exp(V_j^T s)}$$

$$L = -\sum_{v} \log p_{t_v}^S(v) - \sum_{s} \log p_{t_s}^V(s)$$

Non-parametric

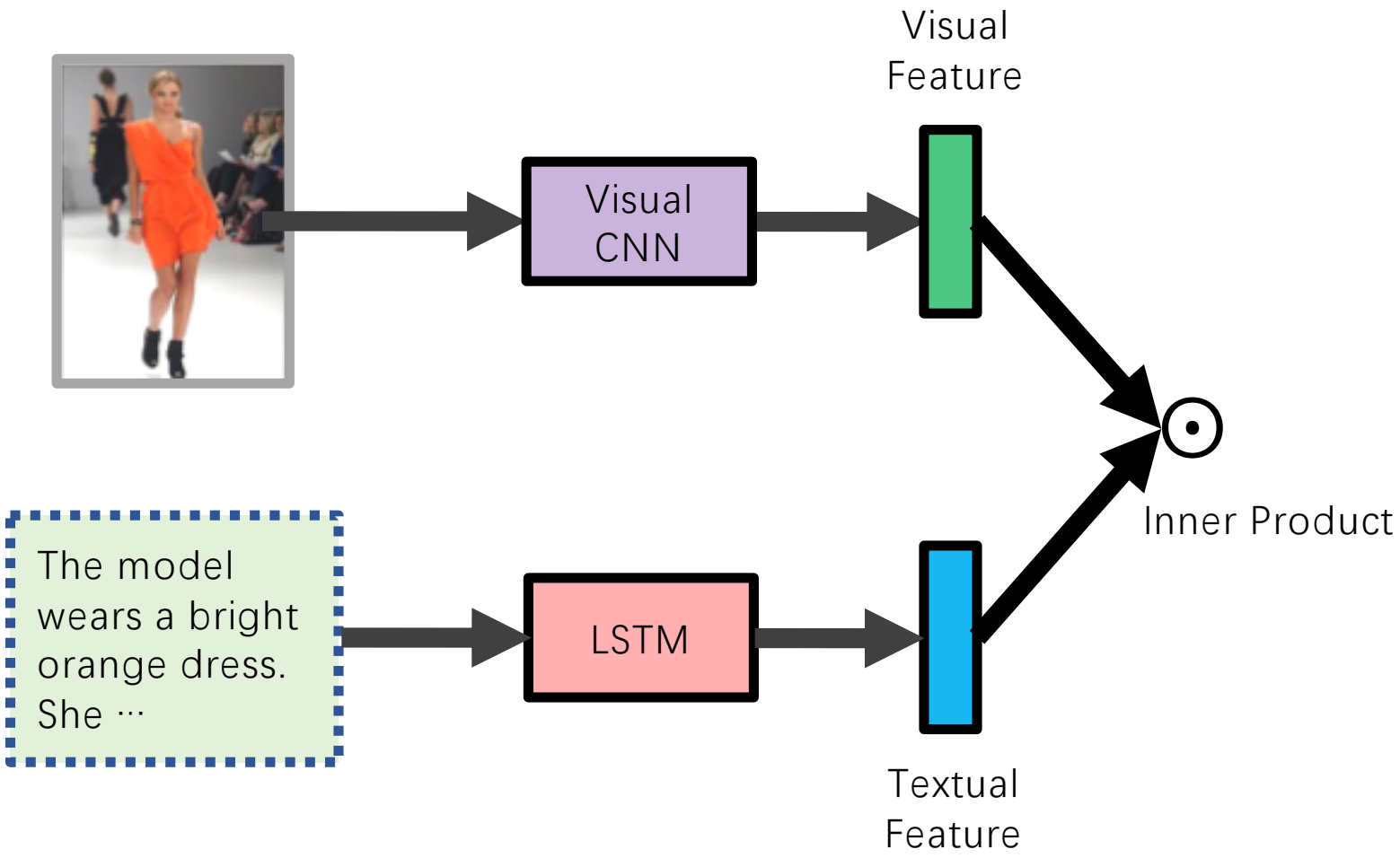
Gradients w.r.t. Features



Minimize distance between same person

Maximize distances among different people

Stage-1 Test Phase



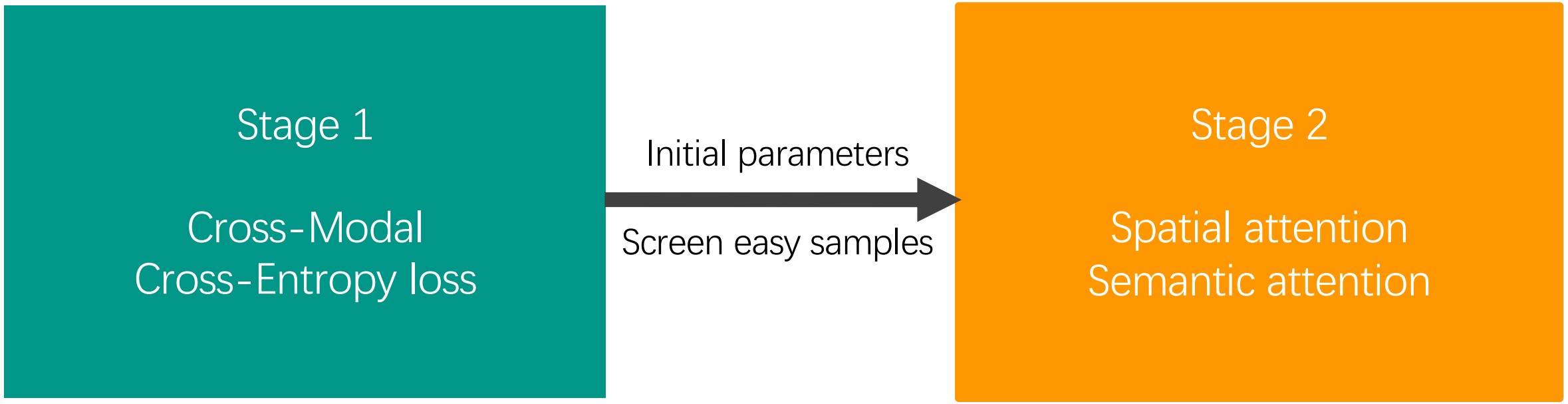
Comparison: CMCE vs. Others

Solutions	Phase	Limitations	CMCE
Pair-wise Loss	Test	N^2 visual-textual pairs	N images + N texts
Triplet Loss	Train	Difficulty in selecting hard negative samples	Hard negative data in each epoch
Classification Loss	Train	Few positives for classifier matrix; different feature embedding space	Non-parametric; Same feature subspace

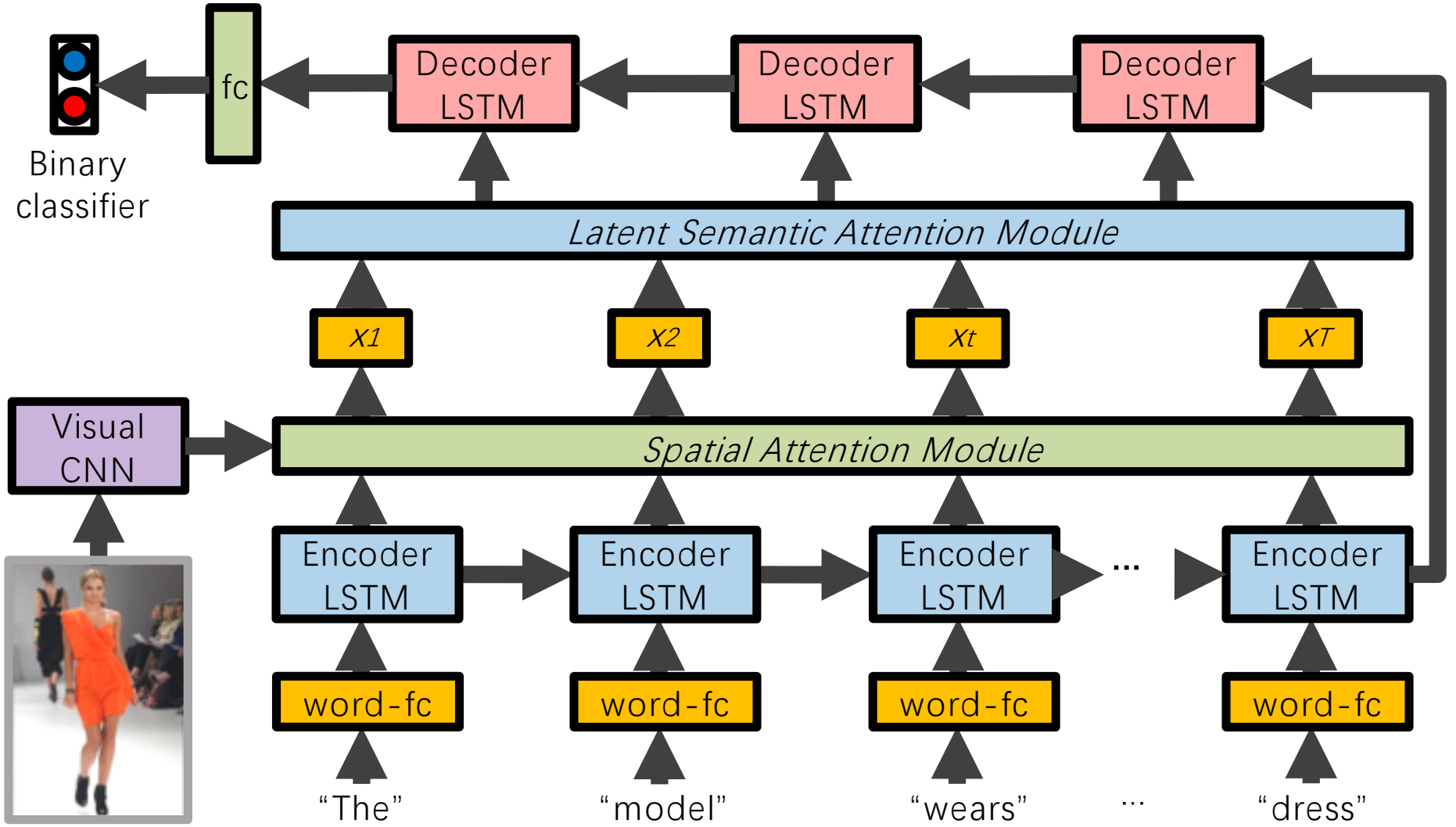
Problems with Stage-1

- Compress the whole sentence into a single vector
- Highly focus on the latest words
- Sensitive to sentence structure variations
  - “The girl who has brown hair is wearing a white dress .”
  - “The girl wears a white dress. She has brown hair.”

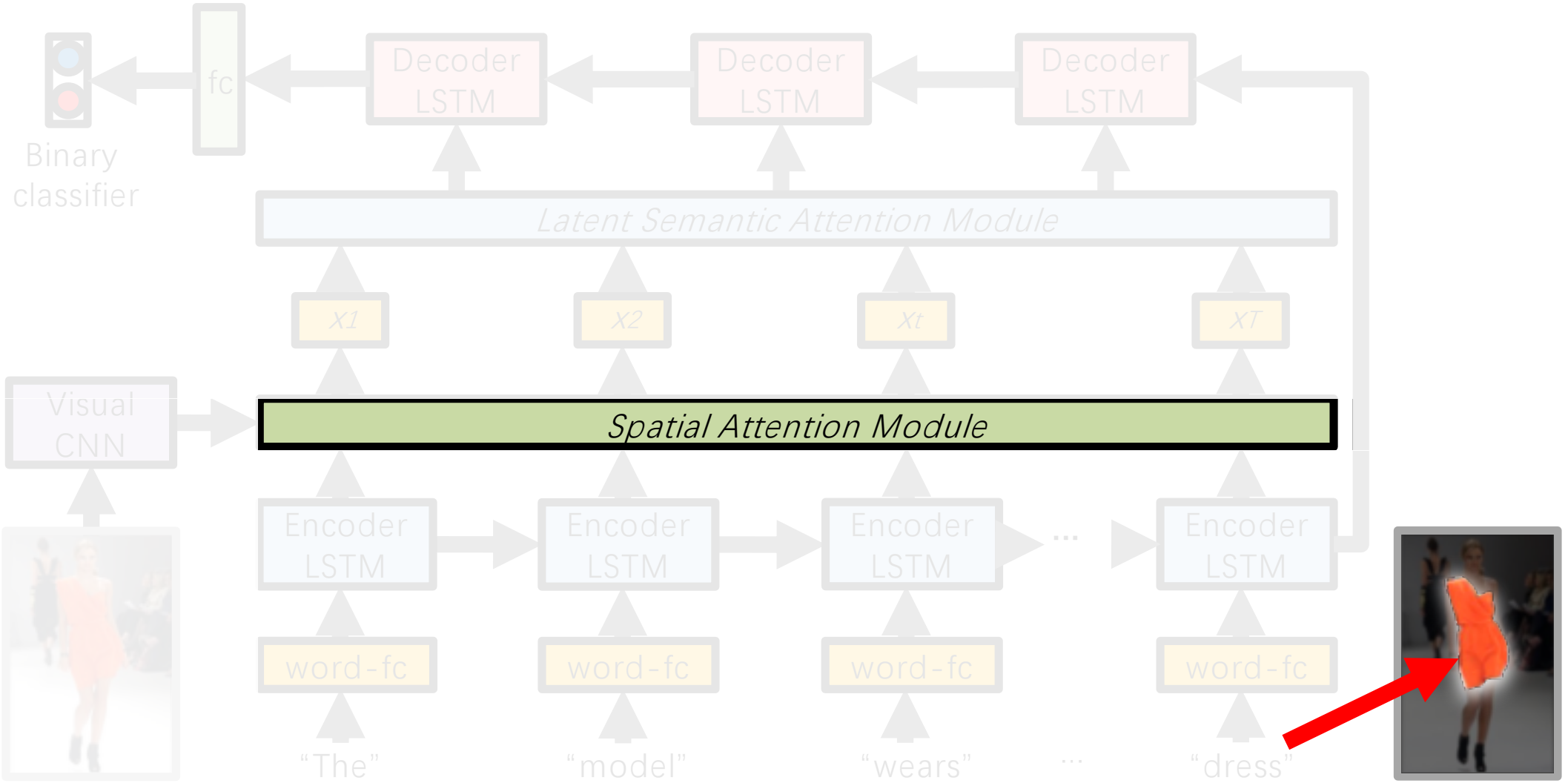
Our framework



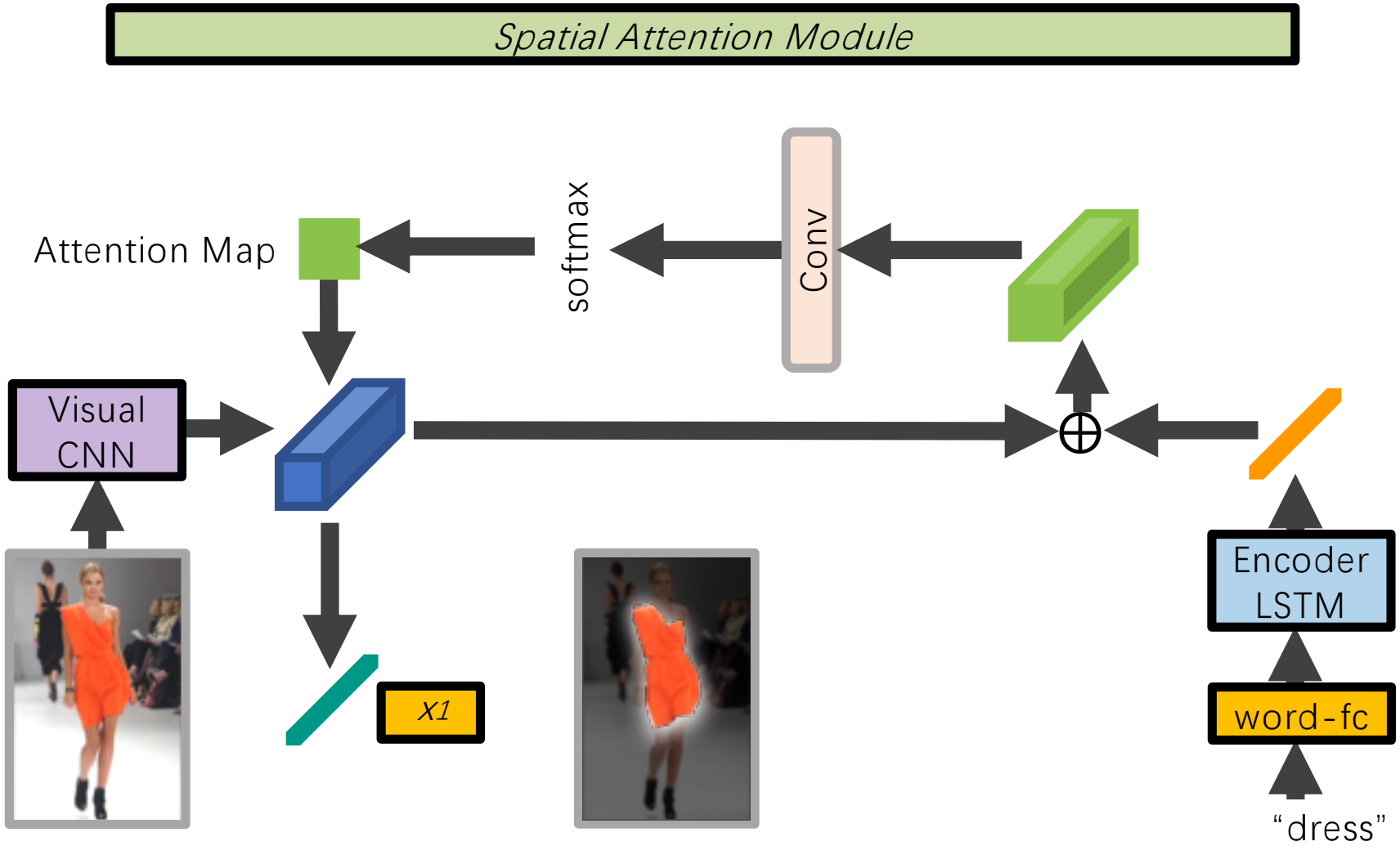
Stage-2 CNN-LSTM with latent co-attention



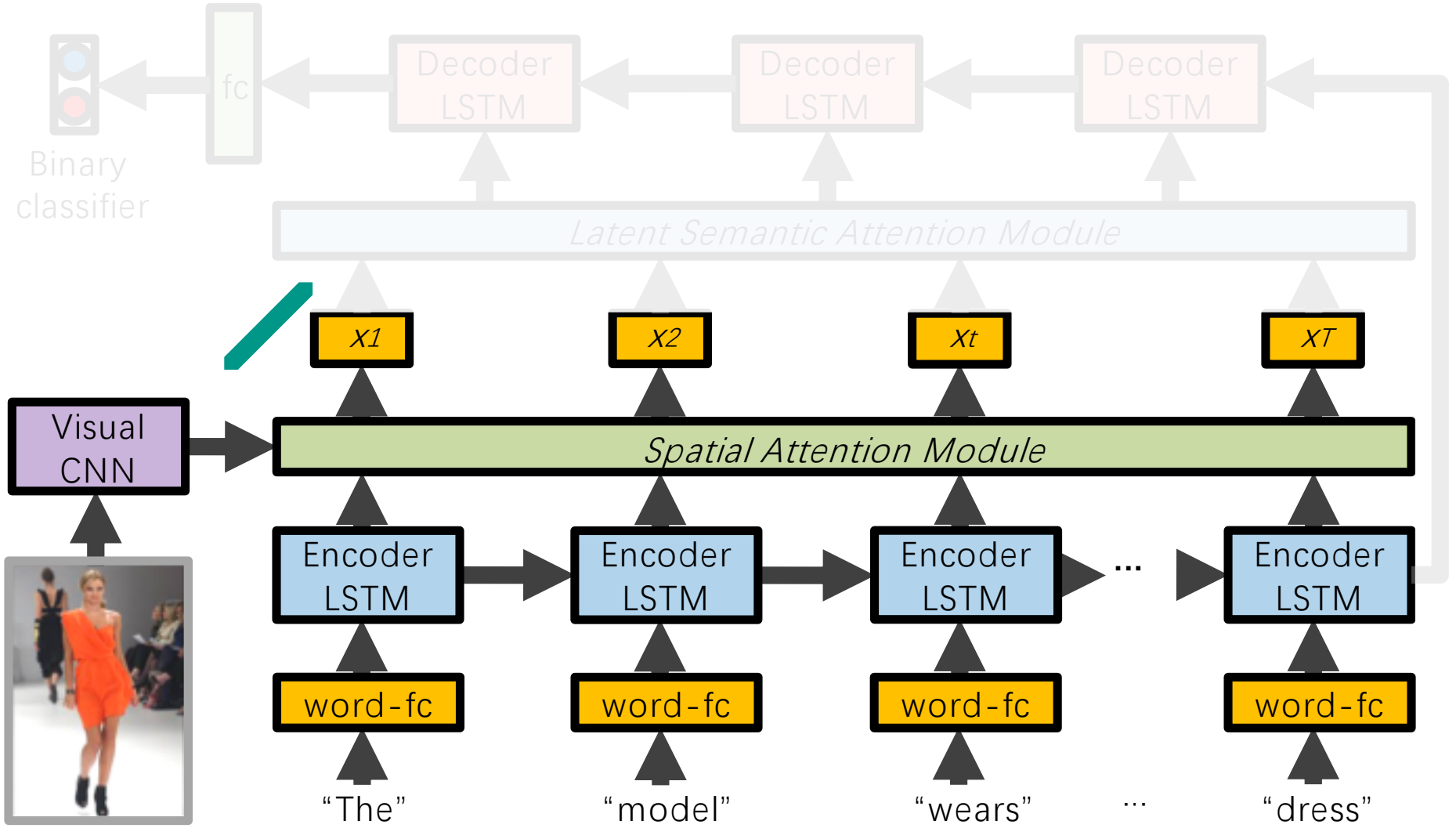
Stage-2 CNN-LSTM with latent co-attention



Spatial Attention



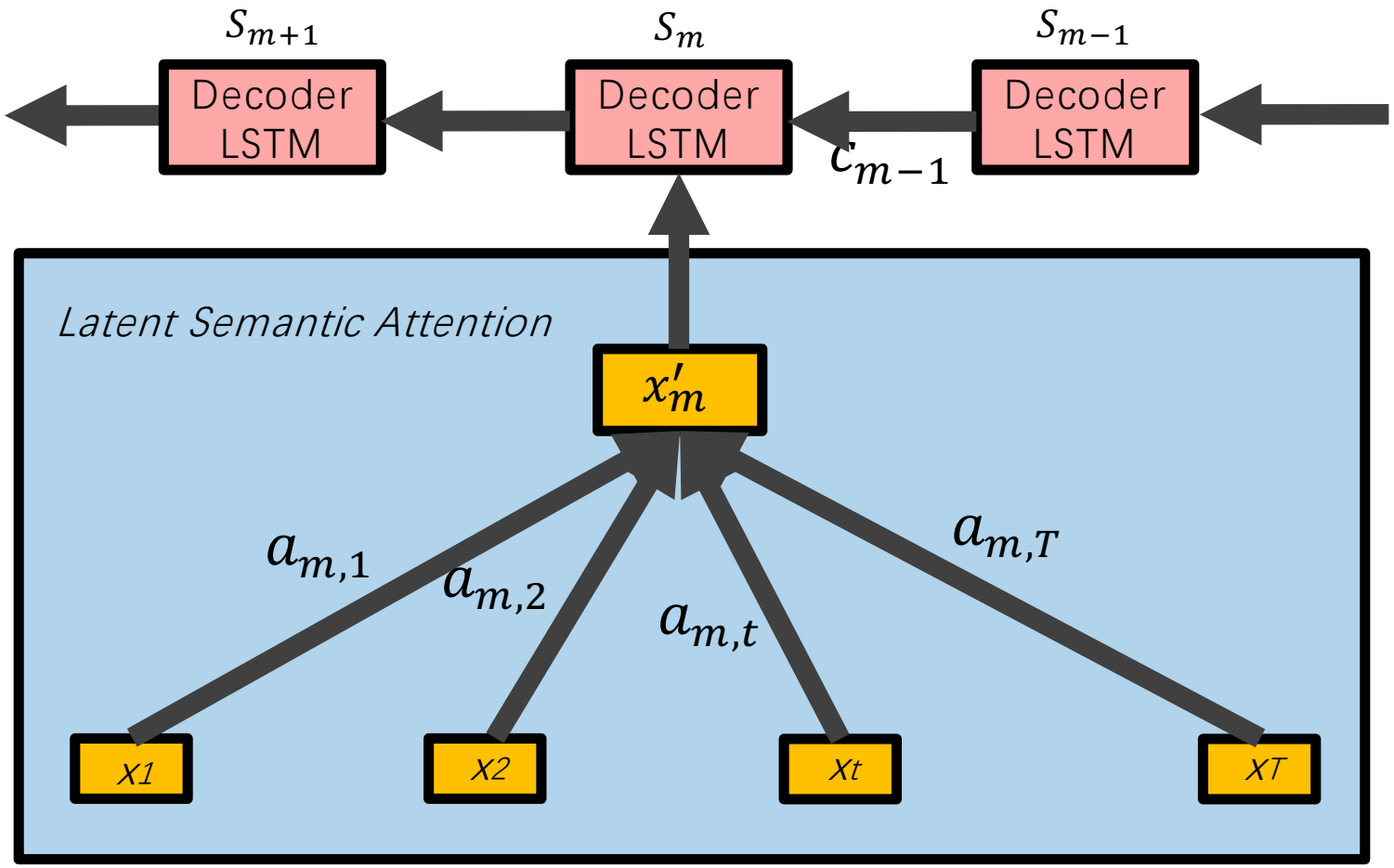
Spatial Attention



Latent Semantic Attention

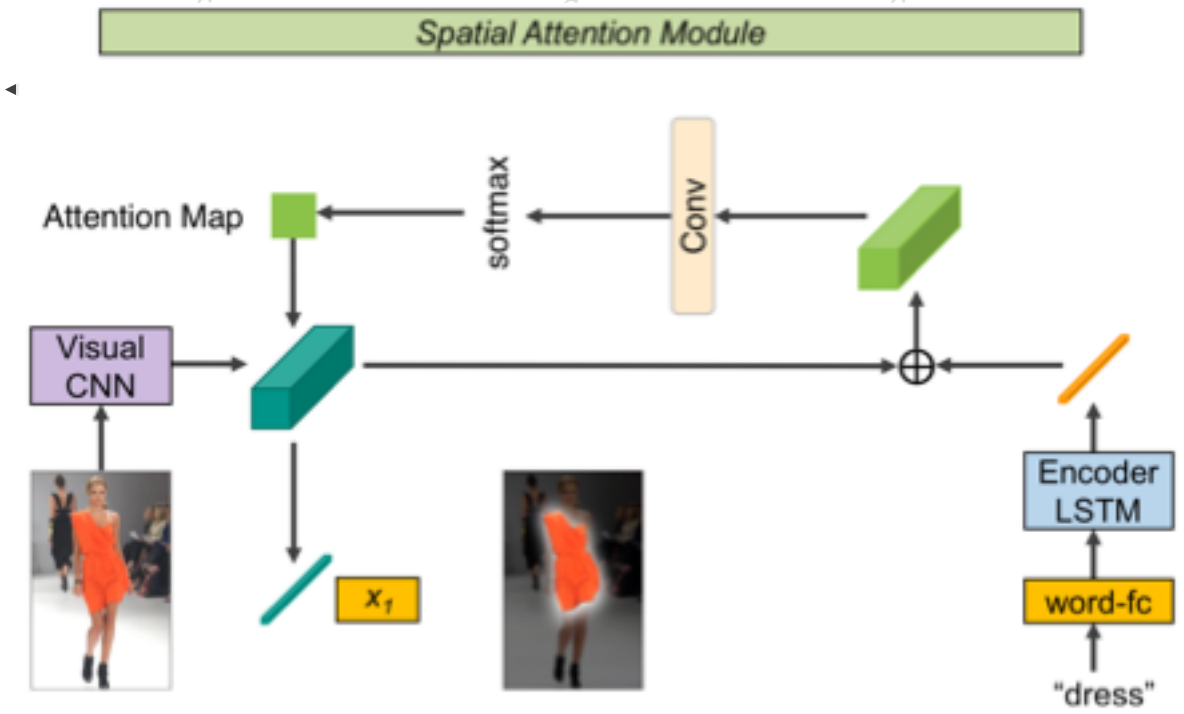
$$x'_m = \sum_{t=1}^{T} a_{m,t} x_t$$

$$a_{m,t} = softmax(f(c_{m-1}, x_t))$$

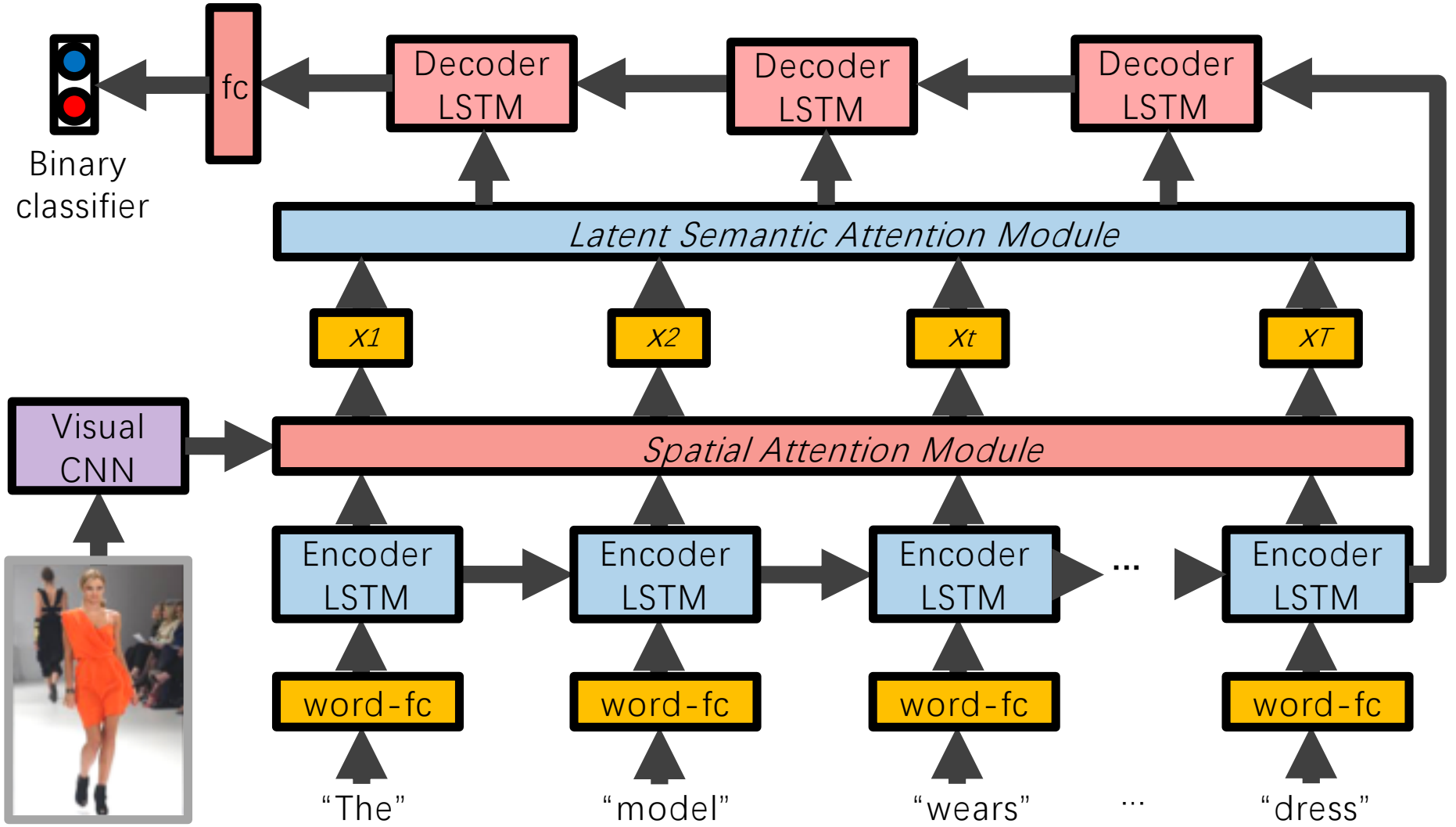


Problems with Stage-1

- Compress the whole sentence into a single vector
- Highly focus on the latest words
- Sensitive to sentence structure variations

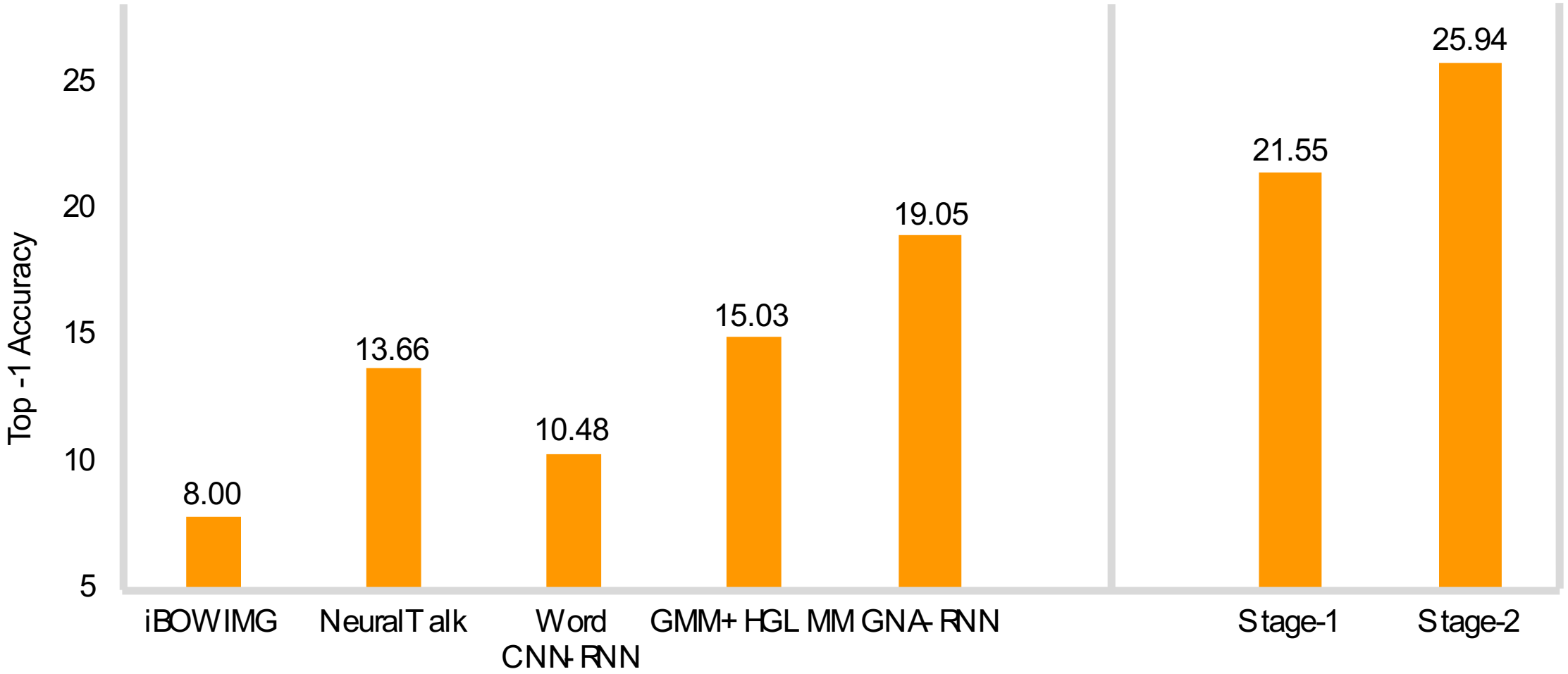


Latent Semantic Attention

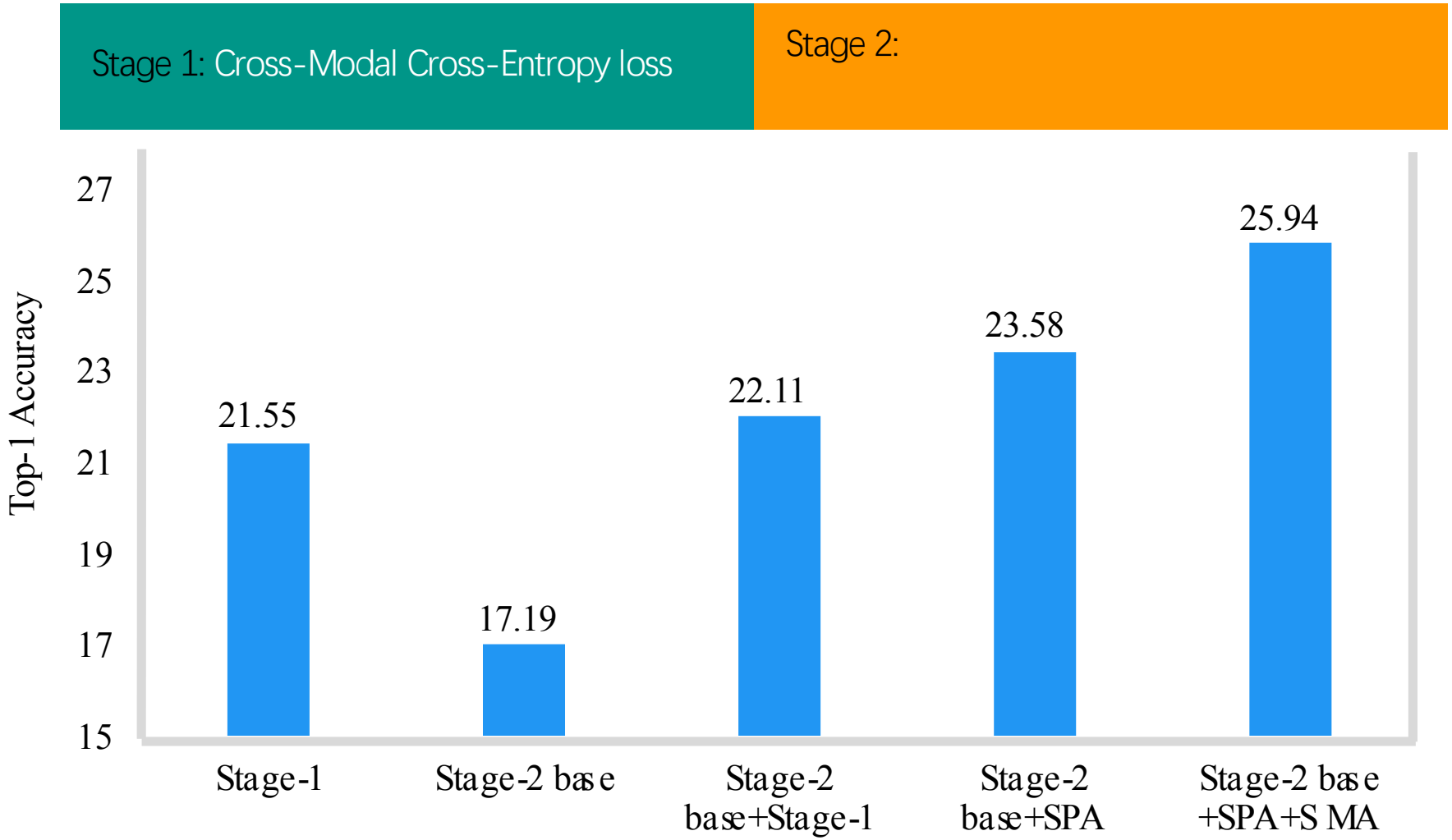


Experiments

Comparisons with different methods



Performance Analysis



Qualitative Results

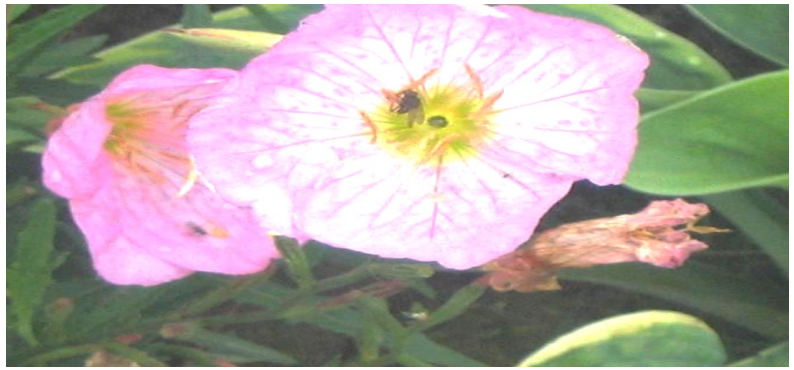
The man is wearing a blue and white striped shirt and black pants. He is also wearing black shoes.

A man is wearing an orange button up shirt, a pair of white pants, and a pink umbrella in his right hand.

Textual-Visual Datasets

CUB Dataset (200 categories)

Flower Dataset (102 categories)



Text-Image Retrieval Results

A long red beaked bird with a white breast and turquoise secondaries.

This bird has a blue head and back with dark secondaries.

These petals are large and purple in color with green in the middle.

The petals on this flower are mostly white with a yellow stamen.

Image-Text Retrieval Results

1. This bird has a black head, short beak, yellow breast and underbelly and has some white mixed with black on its wings.
2. This colorful bird has a bright orange breast belly and tail while being almost black elsewhere.
3. Bright orange bellied and breast bird with black head and wings with gray wing bars.
4. A small sized bird with a vibrant yellow orange breast and underside black head and eyes and black wings with white along the tips.
5. This bird is orange within its underbelly and black around its head and wings and it also has a sharp beak and black eyes.

1. A water swimming bird with brownish body and red eyes.
2. A medium size water fowl with red eyes.
3. A gray and white speckled water fowl with long neck red eye and black wings.
4. This particular bird has a long brown neck and red eyes.
5. This bird has a black beak, orange eyes, a brown crown brown throat, light brown neck and dark brown wings with white secondaries.

1. This flower has a horizontal pedicel and stem with sharply pointed upright orange petals.
2. This flower has a dusty pink pedicel with contrasting upright bright and orange petals with pointed tips.
3. This flower has a lightly multicolored pedicel that holds the upright sharply pointed orange petals.
4. This flower has a white receptacle holding upright long orange pointed petals.
5. This flower has long spiky petals some of which are orange and some of which are blue.

1. This flower has white petals some of which have red lines on them and white filaments with brown anthers.
2. Zygomorphic white flower has burgundy nectar guides on the top two petals and prominent stigma and stamen.
3. This flower has large white petals with dark red stripes leading to the center.
4. The petals of this flower are white and irregular with pink at their base an no prominent stamen or pistil.
5. This has a large white pedal with several pedals having pink accents and several stamen extending out the middle.

Conclusion

- Propose a two-stage strategy for textual-visual matching
- Design a Cross-modal Cross-entropy loss function in Stage-1
- Present a Spatial attention and Semantic attention in Stage-2

Person Search

- **Image** based person search

- **Video** based person search

- **Natural language** based person search

The woman is wearing a long, bright orange gown with a white belt at her waist.

Conclusions

Publications

- Shuang Li*, Tong Xiao*, Bochao Wang, Liang Lin, and Xiaogang Wang, "Joint Detection and Identification Feature Learning for Person Search." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang, "Person Search with Natural Language Description." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang, "Identity-Aware Textual-Visual Matching with Latent Co-attention." In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
- Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, "Learning Feature Pyramids for Human Pose Estimation." In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
- Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang, "Diversity Regularized Spatiotemporal Attention for Video-based Person Re-identification." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

Google Citation

Shuang Li

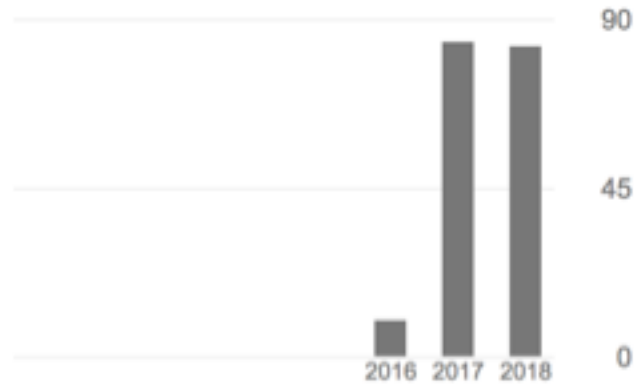
The Chinese University of Hong Kong
Verified email at ee.cuhk.edu.hk - Homepage
Computer Vision Deep Learning

FOLLOW

TITLE	CITED BY	YEAR
Joint detection and identification feature learning for person search S Li, T Xiao, B Wang, L Lin, X Wang Computer Vision and Pattern Recognition (CVPR), 2017	64	2017
End-to-end deep learning for person search S Li, T Xiao, B Wang, L Lin, X Wang arXiv preprint arXiv:1604.01850	36	2016
Person search with natural language description S Li, T Xiao, H Li, B Zhou, D Yue, X Wang Computer Vision and Pattern Recognition (CVPR), 2017	26	2017

Cited by

	All	Since 2013
Citations	180	180
h-index	6	6
i10-index	5	5



Reviewer

- **Journal Reviewer**

  IEEE Transactions on Image Processing (TIP)

  IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)

  IEEE Transactions on Cybernetics

  Computer Vision and Image Understanding (CVIU)

  Image and Vision Computing (IMAVIS)

- **Conference Reviewer**

  IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)

  ACM International Conference on Multimedia (ACMMM 2018)

Sincere Friendship

Thanks to

My Supervisors
Prof. Xiaogang Wang and Prof. Hongsheng Li

Committee members
Prof. Thierry Blu, Prof. Dahua Lin, and Prof. Guofeng Zhang

My dear colleagues
Especially Tong Xiao, Kai Kang, Wei Yang, Jing Shao, Xiao Chu …