THU-PM-269

Improving Cross-Modal Retrieval with Set of Diverse Embeddings

CVPR 2023, Highlight

Dongwon Kim, Namyup Kim, Suha Kwak

Task definition

What is the cross-modal retrieval?

Text-to-image retrieval

Image-to-text retrieval

"Boys wearing helmets carry a bicycle up a ramp at a skate park."

"Small children stand near bicycles at a skate park."

"A group of young children riding bikes and skateboards."

- Cross-modal retrieval: The task of searching for data relevant to a query from a database when the query and database have different modalities (image and text).

Ambiguity problem

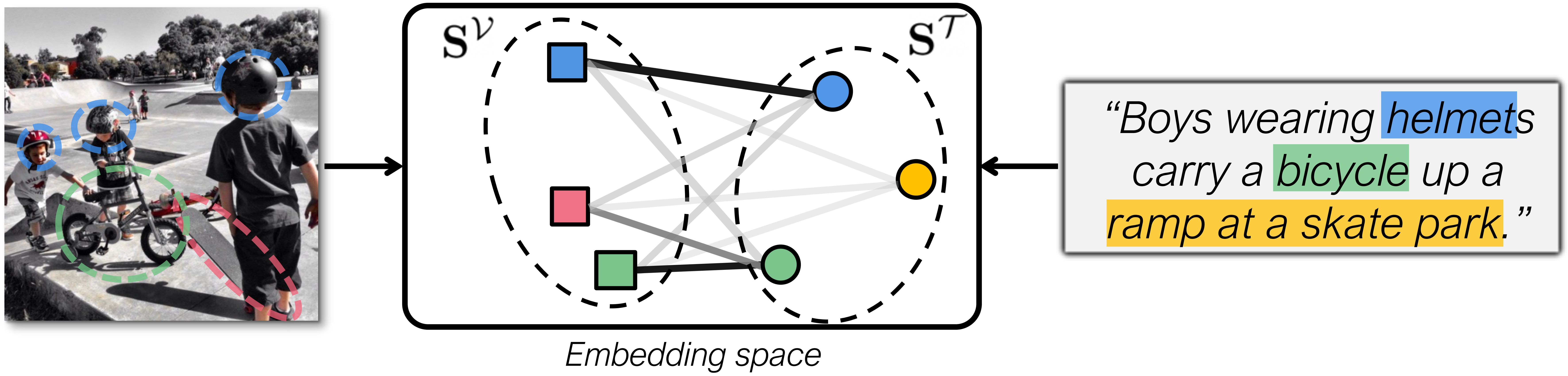
“Boys wearing helmets carry a bicycle up a ramp at a skate park.”

“Small children stand near bicycles at a skate park.”

“A group of young children riding bikes and skateboards.”

- **Image-to-text ambiguity**: An image often contains various contexts, which described with varying captions.
- **Text-to-image ambiguity**: Visual manifestations of a caption vary significantly as captions are highly abstract.

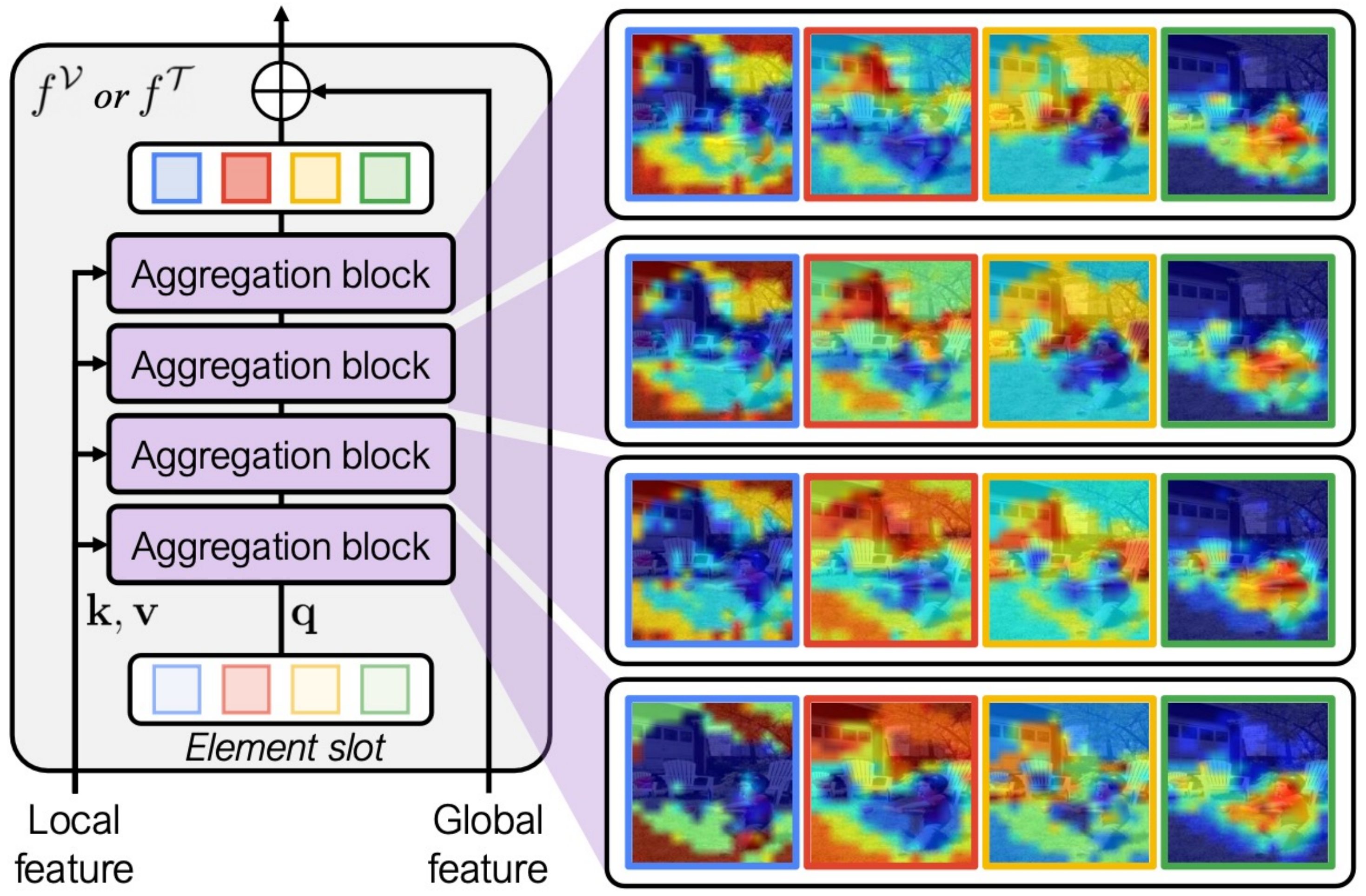
Our method



- Embed the data to a set of diverse embedding vectors, where each elements of the set encodes ***diverse and ambiguous semantics*** of the data.

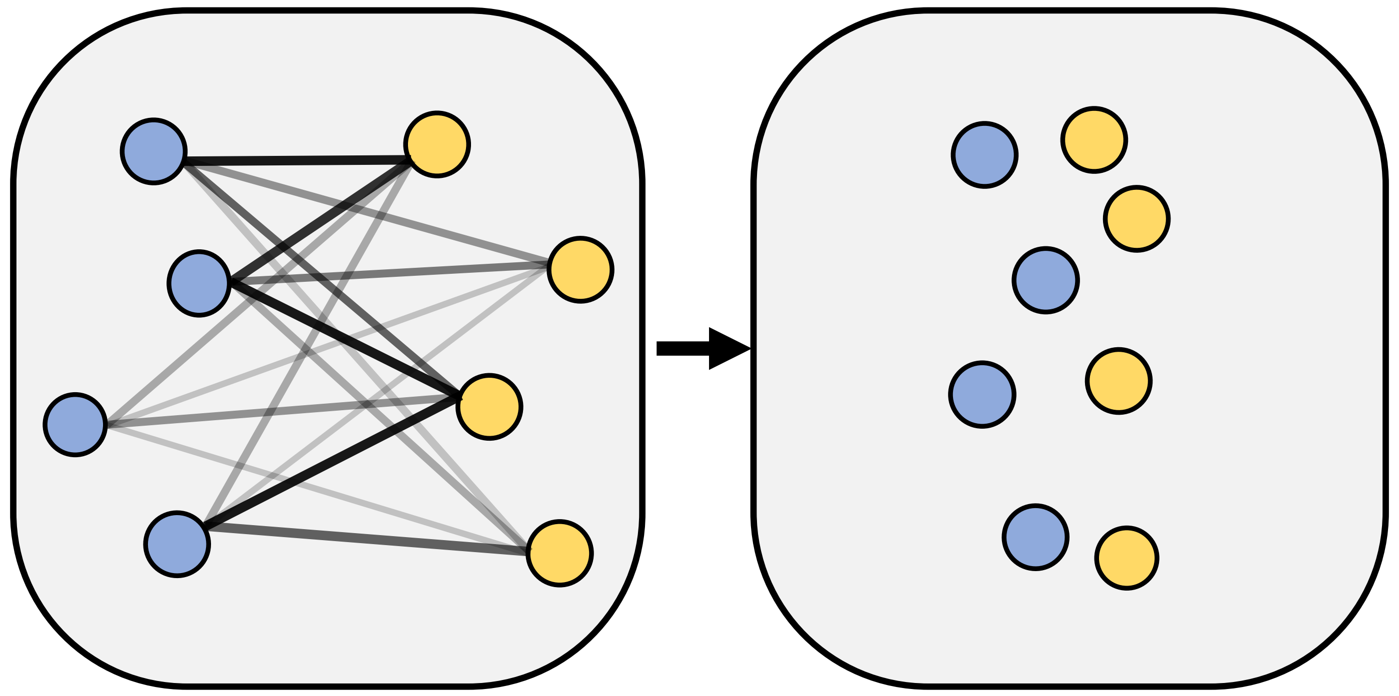
Our method

Set-prediction module



$$\texttt{attn} = \texttt{softmax}\left(\frac{1}{\sqrt{D}} k(\texttt{inputs}) \cdot q(\texttt{slots})^T, \texttt{axis=`slots'}\right)$$

Smooth-Chamfer similarity



$$s_{\mathrm{SC}}\left(\mathbf{S}_1, \mathbf{S}_2\right) = \frac{1}{2\alpha\left|\mathbf{S}_1\right|} \sum_{x \in \mathbf{S}_1} \underset{y \in \mathbf{S}_2}{\mathrm{LSE}}(\alpha c(x, y)) + \frac{1}{2\alpha\left|\mathbf{S}_2\right|} \sum_{y \in \mathbf{S}_2} \underset{x \in \mathbf{S}_1}{\mathrm{LSE}}(\alpha c(x, y))$$

[1] Object-centric learning with slot attention, NeurIPS, 2020.

Results

		1K Test Images							5K Test Images						
		Image-to-Text			Text-to-Image				Image-to-Text			Text-to-Image			
Method	CA	R@1	R@5	R@10	R@1	R@5	R@10	RSUM	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
ResNet-152 + Bi-GRU															
VSE++ [17]	✗	64.6	90.0	95.7	52.0	84.3	92.0	478.6	41.3	71.1	81.2	30.3	59.4	72.4	355.7
PVSE [45]	✗	69.2	91.6	96.6	55.2	86.5	93.7	492.8	45.2	74.3	84.5	32.4	63.0	75.0	374.4
PCME [10]	✗	68.8	-	-	54.6	-	-	-	44.2	-	-	31.9	-	-	-
Ours	✗	70.3	91.5	96.3	56.0	85.8	93.3	**493.2**	47.2	74.8	84.1	33.8	63.1	74.7	**377.7**
Faster R-CNN + Bi-GRU															
SCAN† [30]	✓	72.7	94.8	98.4	58.8	88.4	94.8	507.9	50.4	82.2	90.0	38.6	69.3	80.4	410.9
VSRN† [31]	✗	76.2	94.8	98.2	62.8	89.7	95.1	516.8	53.0	81.1	89.4	40.5	70.6	81.1	415.7
CAAN [53]	✓	75.5	95.4	98.5	61.3	89.7	95.2	515.6	52.5	83.3	90.9	41.2	70.3	82.9	421.1
IMRAM† [6]	✓	76.7	95.6	98.5	61.7	89.1	95.0	516.6	53.7	83.2	91.0	39.7	69.1	79.8	416.5
SGRAF† [14]	✓	79.6	96.2	98.5	63.2	90.7	96.1	524.3	57.8	-	91.6	41.9	-	81.3	-
VSE$_\infty$ [27]	✗	78.5	96.0	98.7	61.7	90.3	95.6	520.8	56.6	83.6	91.4	39.3	69.9	81.1	421.9
NAAF† [52]	✓	80.5	96.5	98.8	64.1	90.7	96.5	527.2	58.9	85.2	92.0	42.5	70.9	81.4	430.9
Ours	✗	79.8	96.2	98.6	63.6	90.7	95.7	524.6	58.8	84.9	91.5	41.1	72.0	82.4	430.7
Ours†	✗	80.6	96.3	98.8	64.7	91.4	96.2	**528.0**	60.4	86.2	92.4	42.6	73.1	83.1	**437.8**
ResNeXt-101 + BERT															
VSE$_\infty$ [27]	✗	84.5	98.1	99.4	72.0	93.9	97.5	545.4	66.4	89.3	94.6	51.6	79.3	87.6	468.9
VSE$_\infty$† [27]	✗	85.6	98.0	99.4	73.1	94.3	97.7	548.1	68.1	90.2	95.2	52.7	80.2	88.3	474.8
Ours	✗	86.3	97.8	99.4	72.4	94.0	97.6	547.5	69.1	90.7	95.6	52.1	79.6	87.8	474.9
Ours†	✗	86.6	98.2	99.4	73.4	94.5	97.8	**549.9**	71.0	91.8	96.3	53.4	80.9	88.6	**482.0**

Ambiguity problem

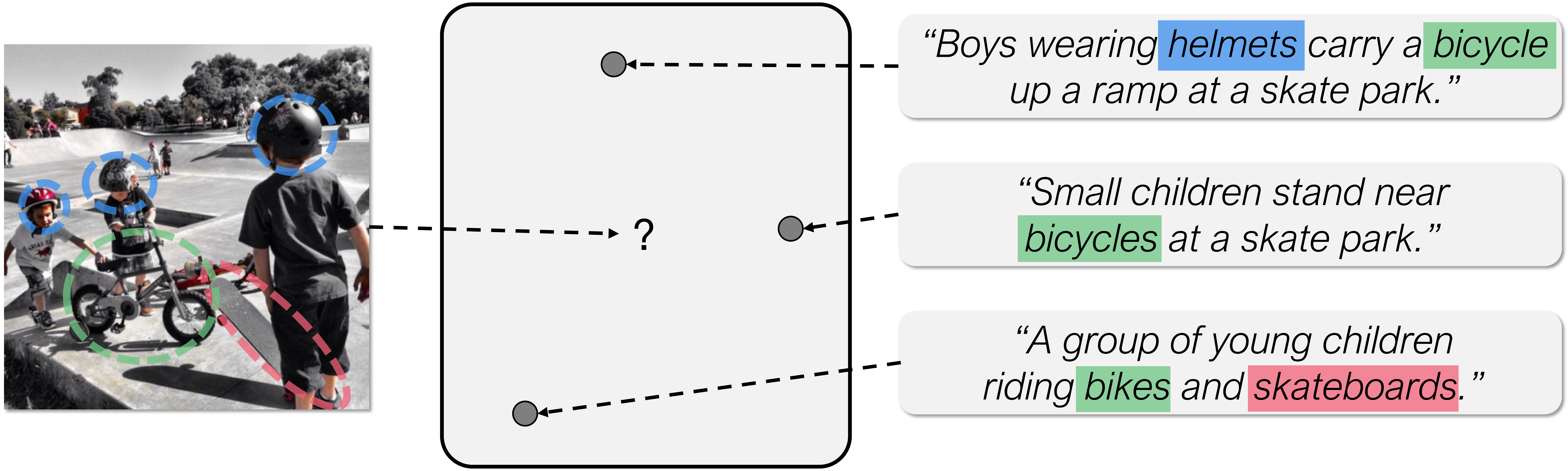
"Boys wearing helmets carry a bicycle up a ramp at a skate park."

"Small children stand near bicycles at a skate park."

"A group of young children riding bikes and skateboards."

- **Image-to-text ambiguity**: An image often contains various contexts, which described with varying captions.
- **Text-to-image ambiguity**: Visual manifestations of a caption vary significantly as captions are highly abstract.

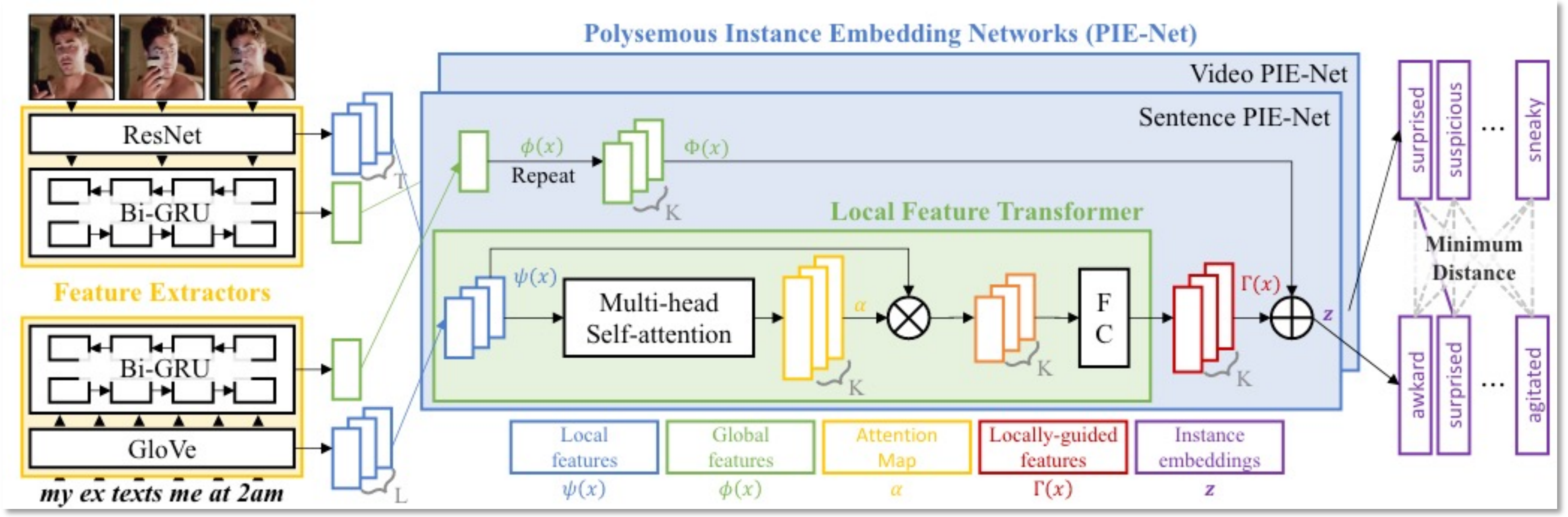
Ambiguity problem

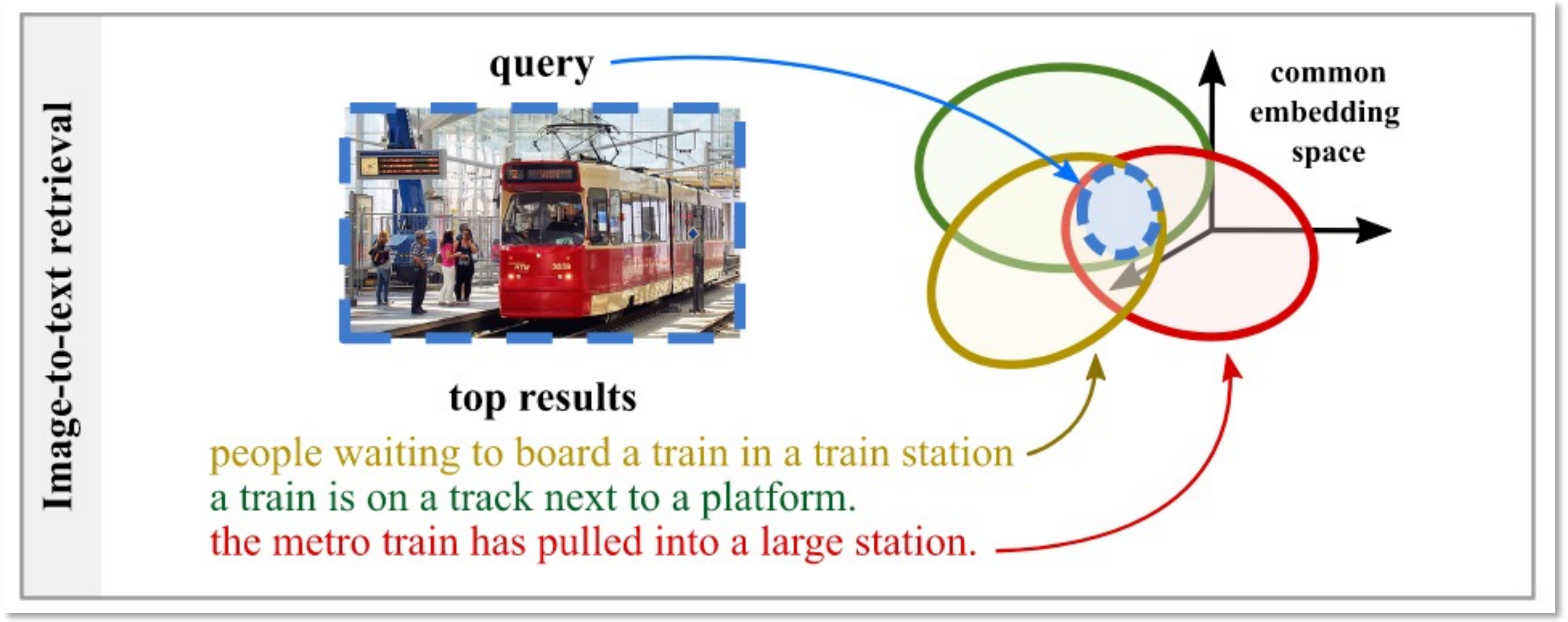


- Conventional embedding models do not resolve the ambiguity problem since they represent a sample as a single embedding vector.

Previous work on set-based embedding

PVSE[1] & PCME[2]





PVSE

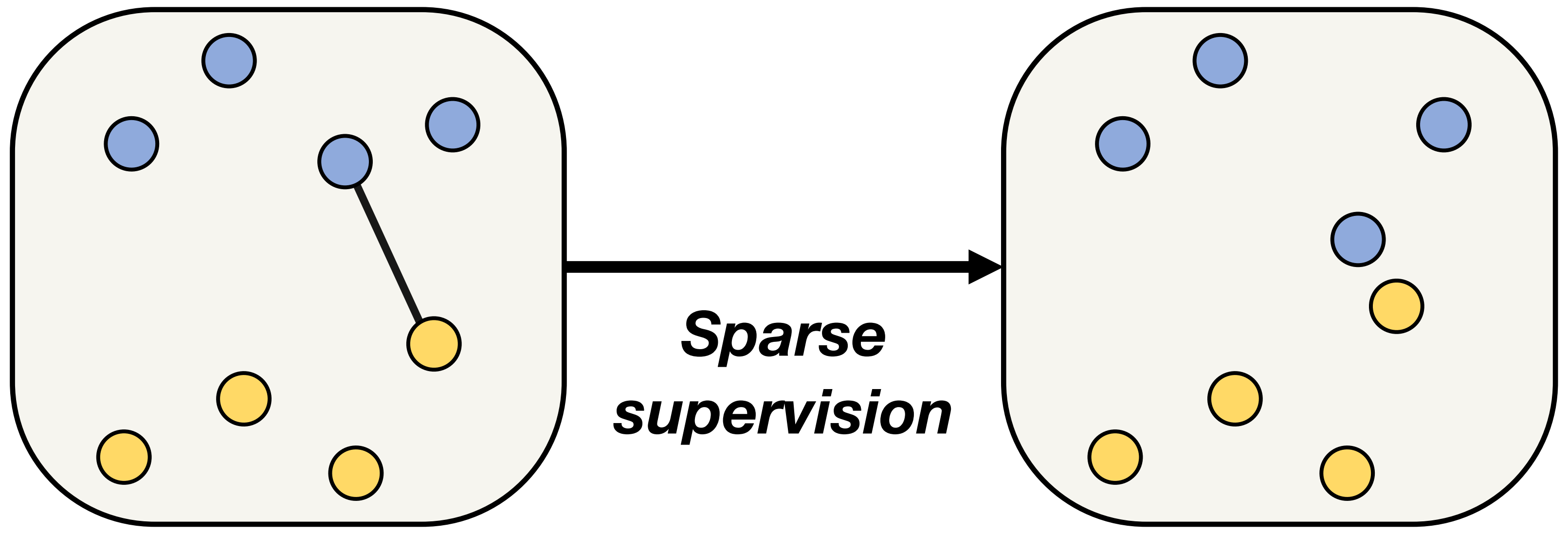
- Represent each sample as a set of embedding vectors

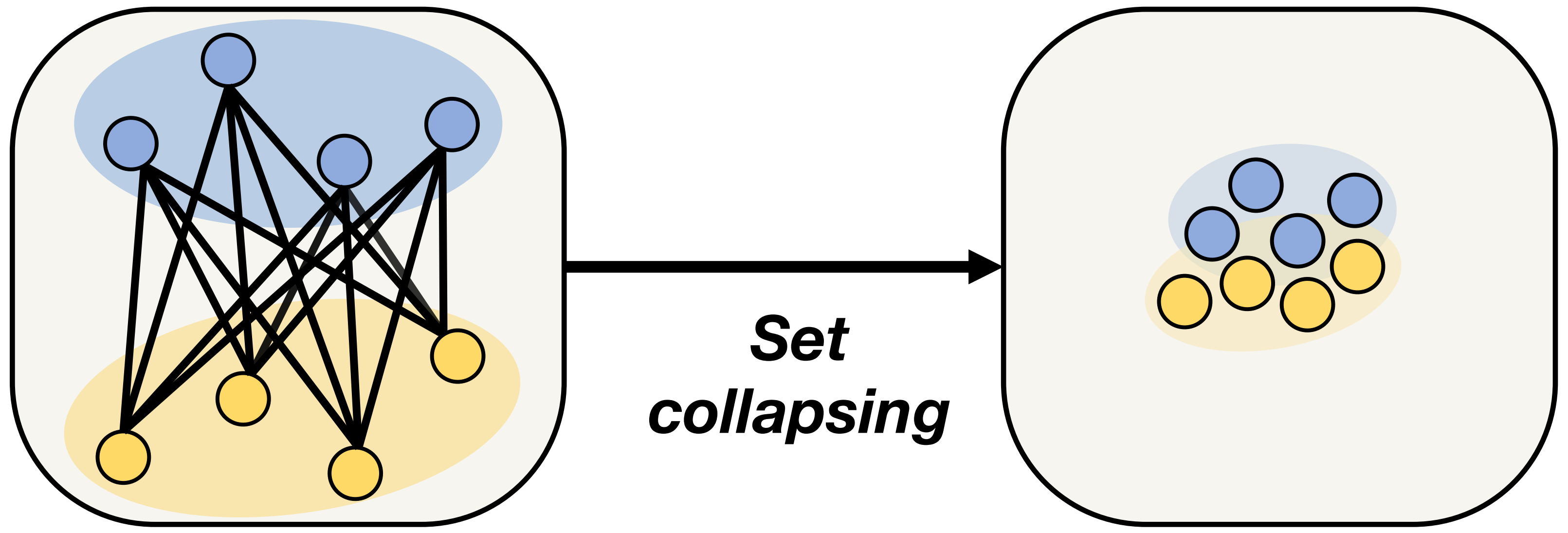
PCME

- Utilize probabilistic embedding where each sample is represented as a set of vectors sampled from a normal distribution

[1] Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval, CVPR, 2019.
[2] Probabilistic Embeddings for Cross-Modal Retrieval, CVPR, 2021.

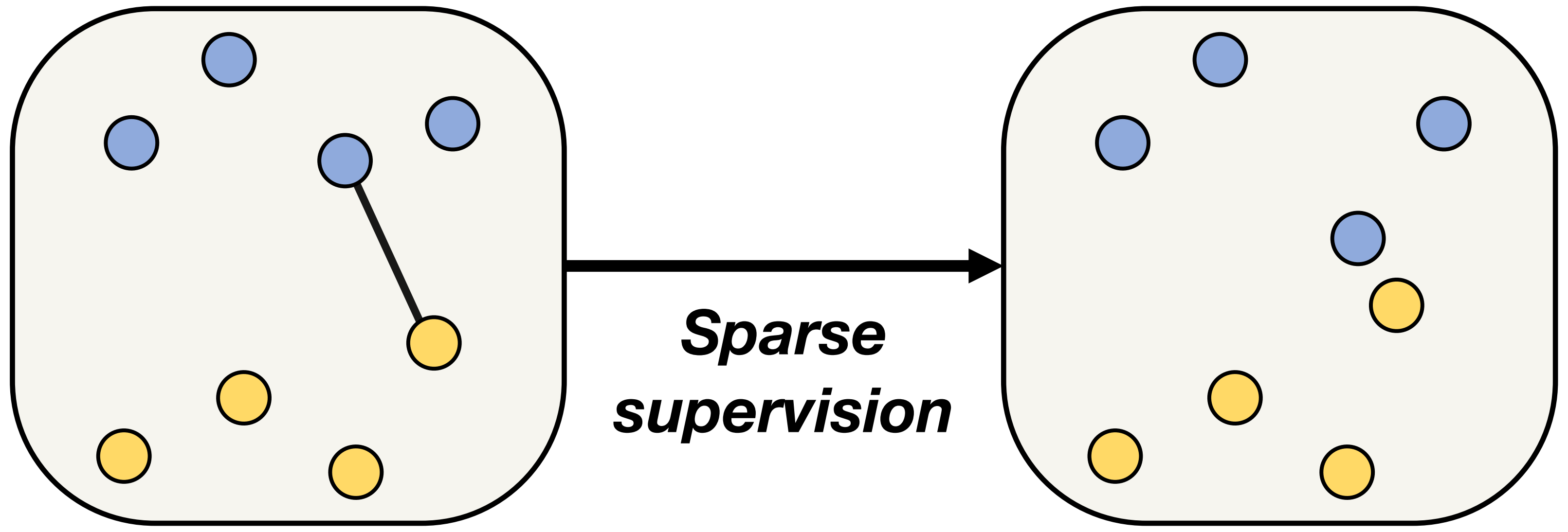
Drawbacks of set-based embedding

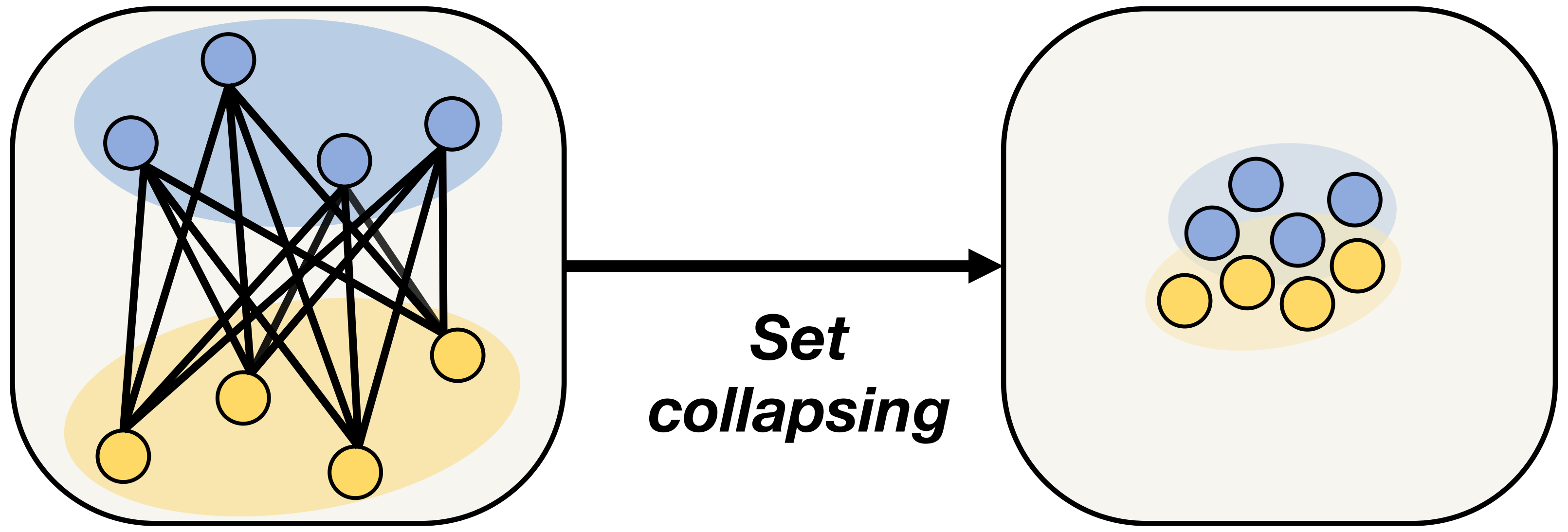




- ***Sparse supervision***→ An embedding set most of whose elements remain untrained.
- ***Set collapsing***→ An embedding set with a small variance which does not encode sufficient ambiguity.

Drawbacks of set-based embedding





- ***Sparse supervision***→ An embedding set most of whose elements remain untrained.
- ***Set collapsing***→ An embedding set with a small variance which does not encode sufficient ambiguity.

Similarity function used for train & eval

- Similarity functions used for training & eval in previous work **do not consider the ambiguity of the data.**

→ ***Sparse supervision, Set collapsing***

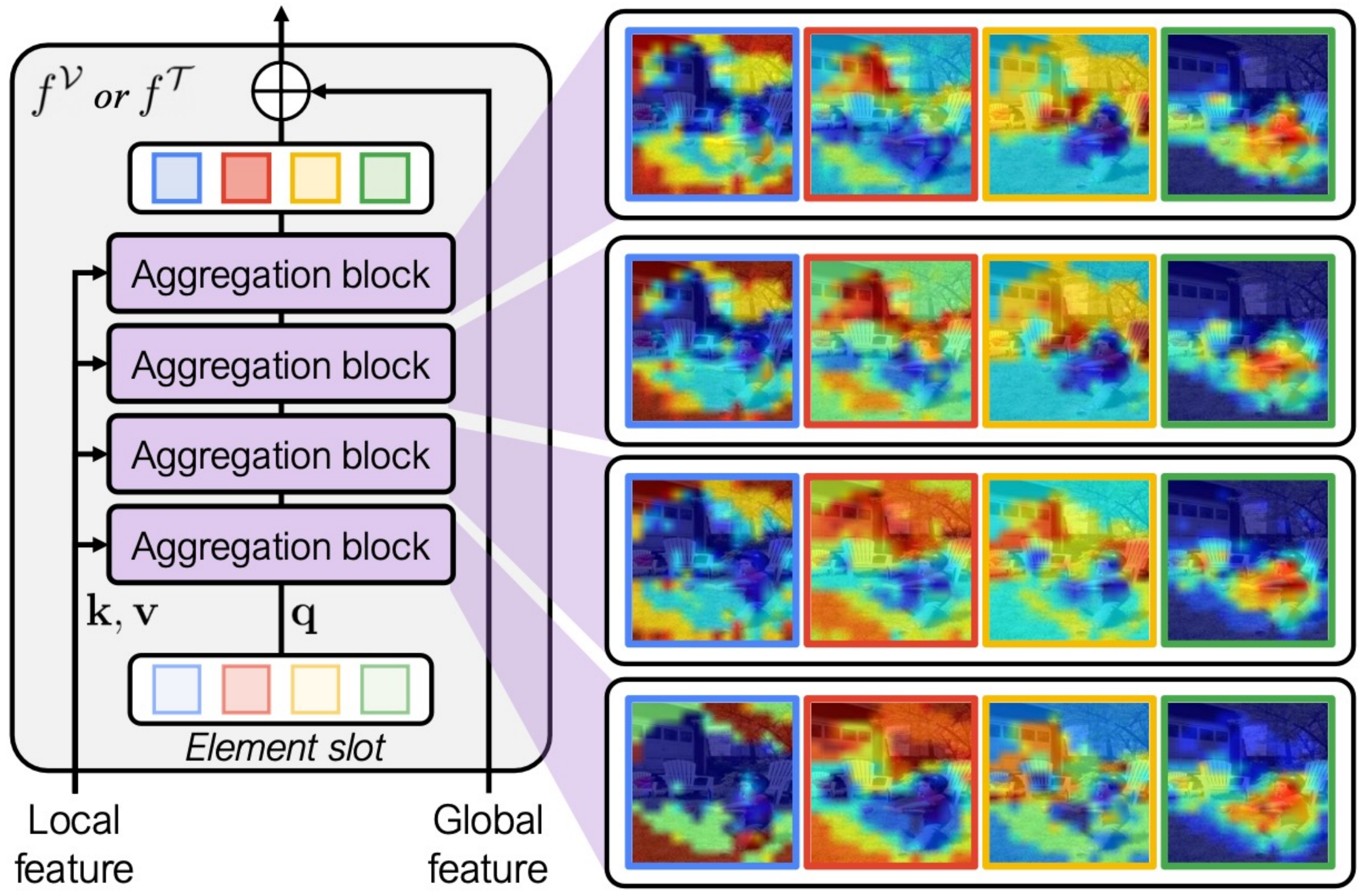
Model architecture for embedding set

- Self-attention modules used for set prediction in the previous work do not explicitly consider **disentanglement between set elements**.

→ ***Set collapsing***

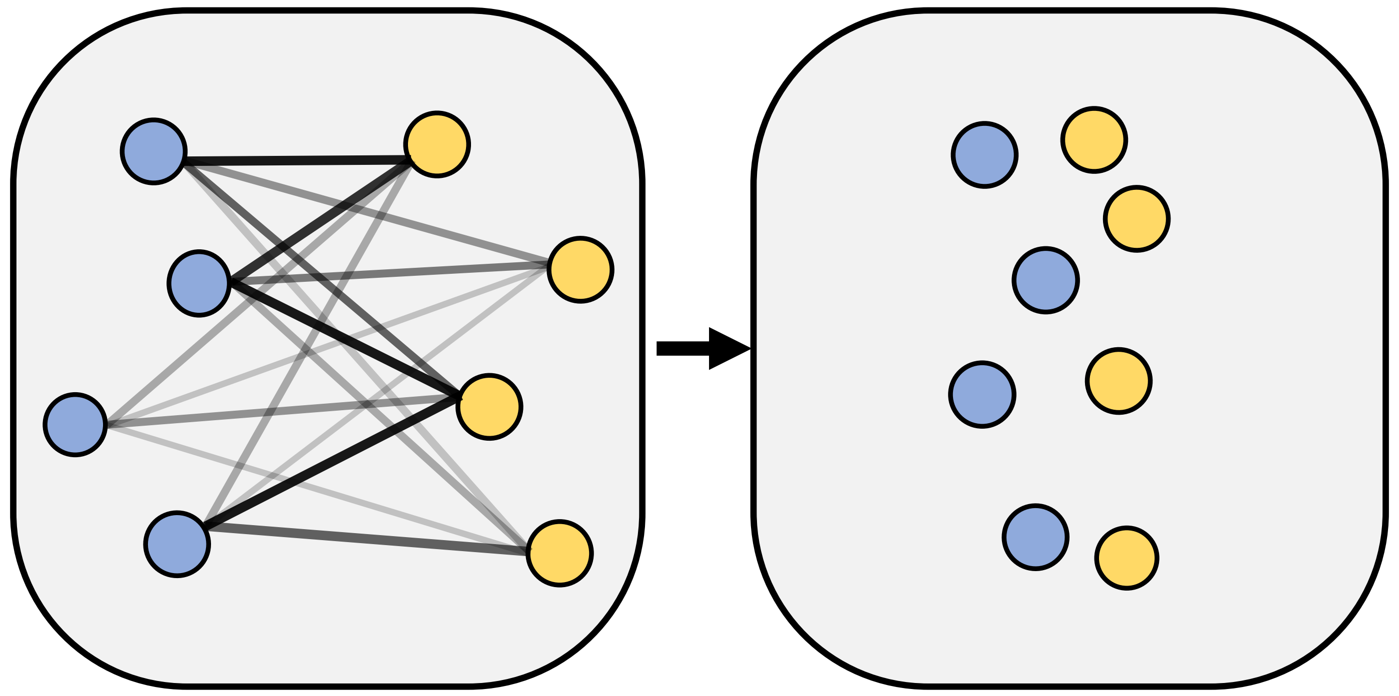
Our method

Set-prediction module



$$\texttt{attn} = \texttt{softmax}\left(\frac{1}{\sqrt{D}} k(\texttt{inputs}) \cdot q(\texttt{slots})^T, \texttt{axis=`slots'}\right)$$

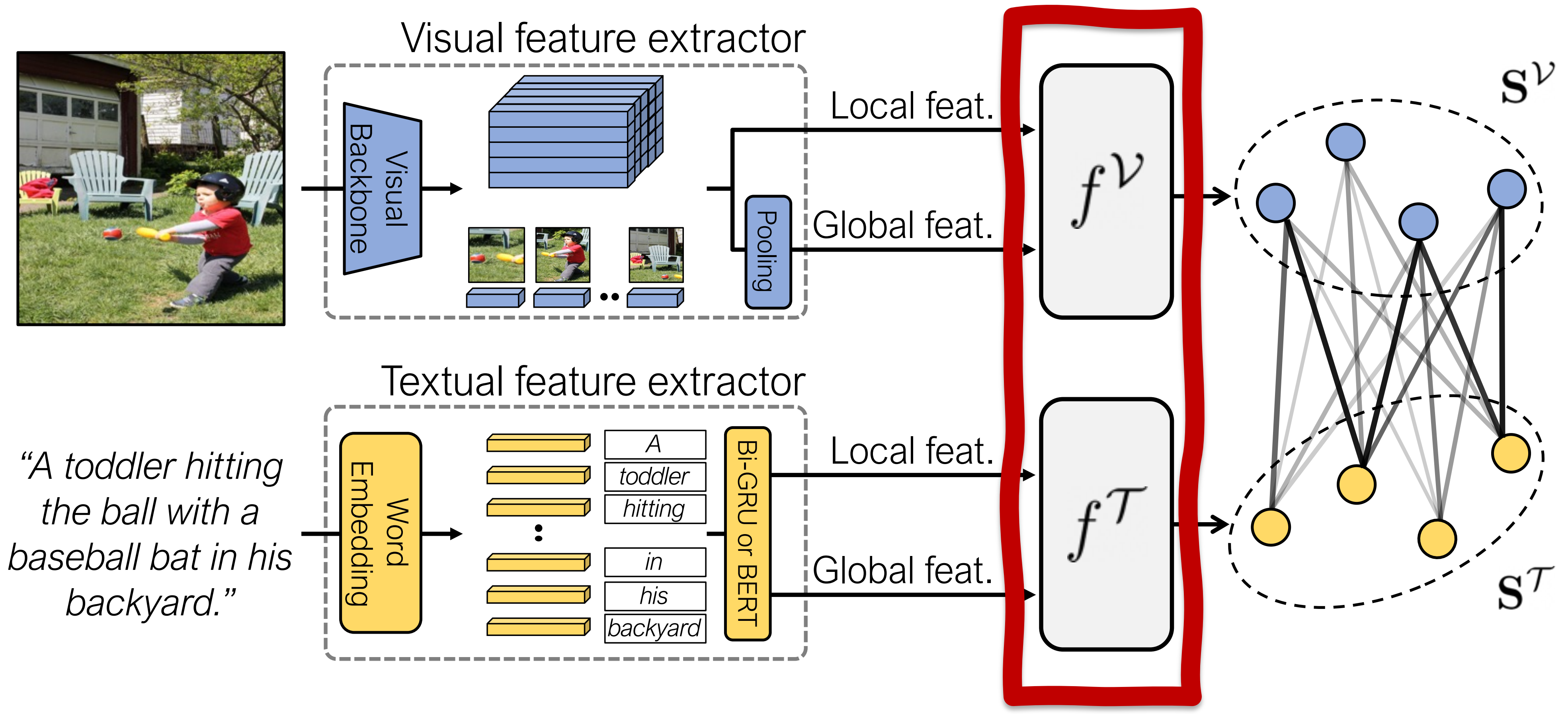
Smooth-Chamfer similarity

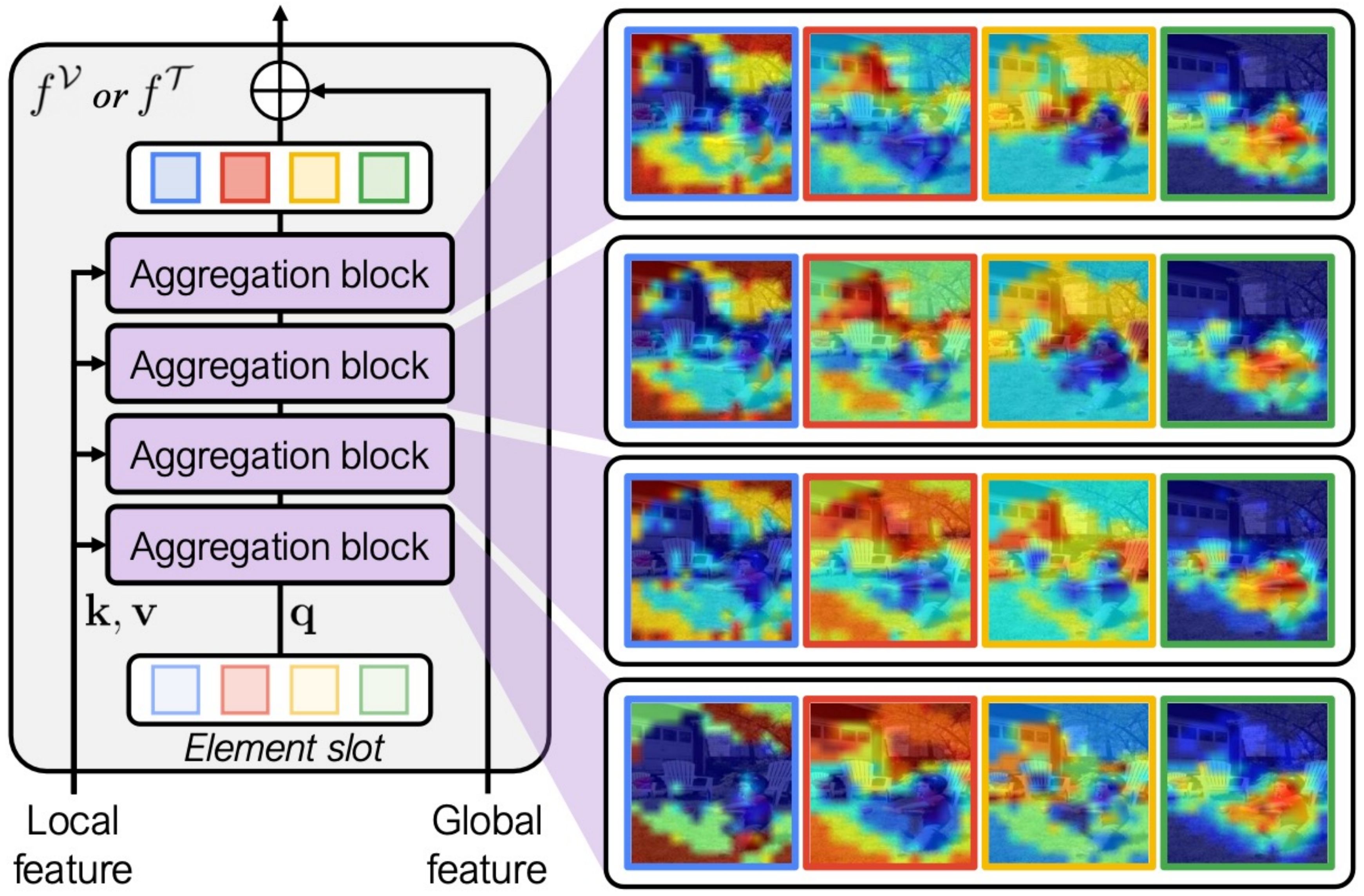


$$s_{\mathrm{SC}}(\mathbf{S}_1, \mathbf{S}_2) = \frac{1}{2\alpha |\mathbf{S}_1|} \sum_{x \in \mathbf{S}_1} \underset{y \in \mathbf{S}_2}{\mathrm{LSE}}(\alpha c(x, y)) + \frac{1}{2\alpha |\mathbf{S}_2|} \sum_{y \in \mathbf{S}_2} \underset{x \in \mathbf{S}_1}{\mathrm{LSE}}(\alpha c(x, y))$$

Overall architecture

2. Set-prediction module





$$\texttt{attn} = \texttt{softmax}\left(\frac{1}{\sqrt{D}} k(\texttt{inputs}) \cdot q(\texttt{slots})^T, \texttt{axis='slots'}\right)$$

Slot-attn[3] based attention scheme (**Ours**)

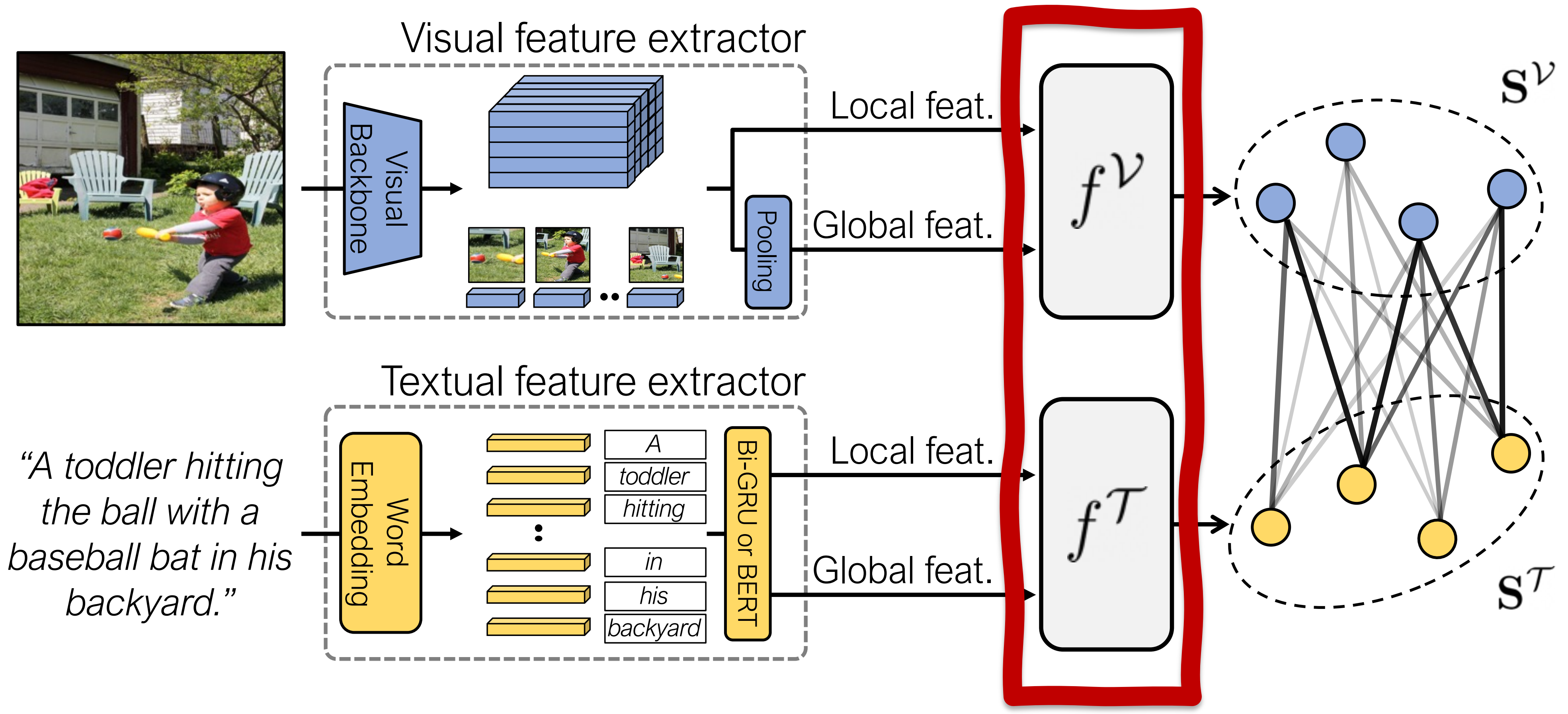
$$\texttt{attn} = \texttt{softmax}\left(\frac{1}{\sqrt{D}} k(\texttt{inputs}) \cdot q(\texttt{slots})^T, \texttt{axis='inputs'}\right)$$

Conventional transformer attention scheme

[3] Object-centric learning with slot attention, NeurIPS, 2020.

Overall architecture

2. Set-prediction module



$$\texttt{attn} = \texttt{softmax}\left(\frac{1}{\sqrt{D}}k(\texttt{inputs})\cdot q(\texttt{slots})^T, \texttt{axis='slots'}\right)$$

Slot-attn [3] based attention scheme (**Ours**)

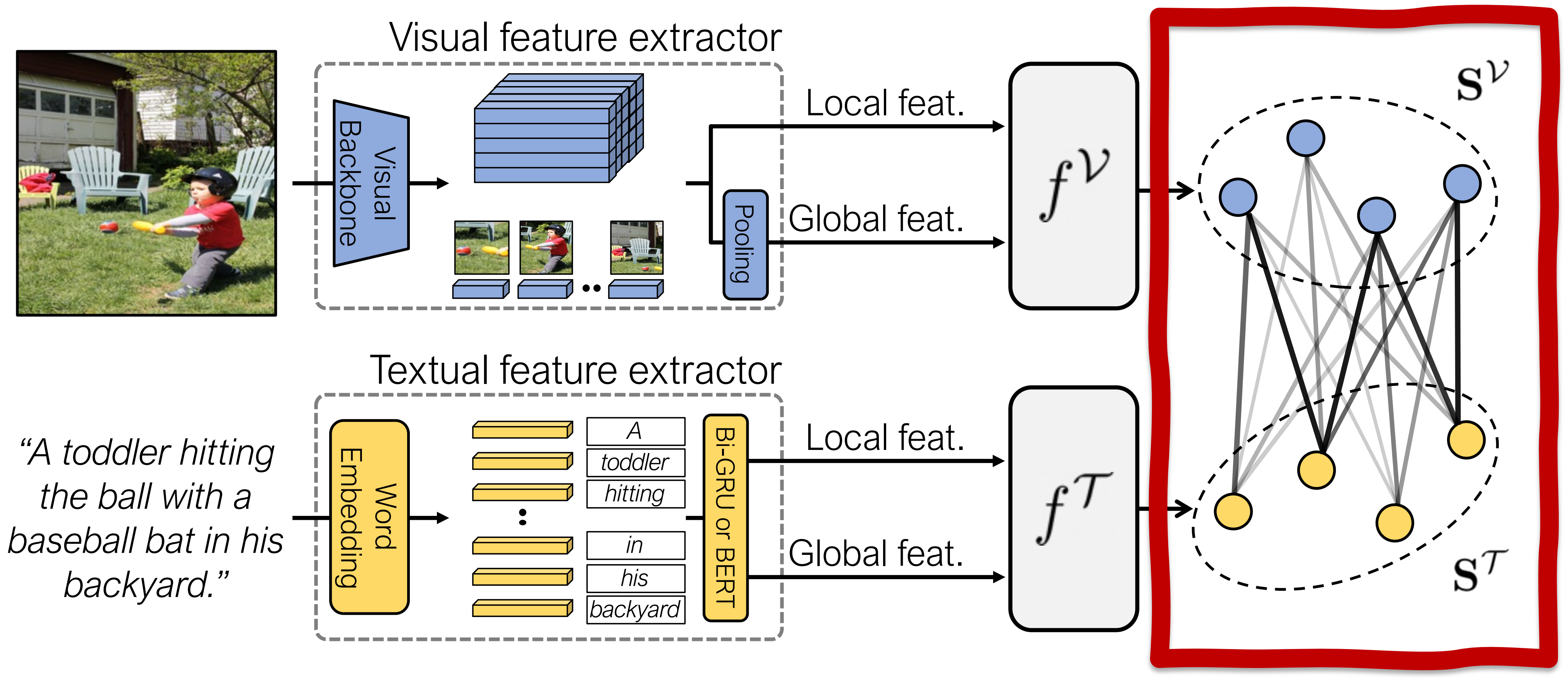
$$\texttt{attn} = \texttt{softmax}\left(\frac{1}{\sqrt{D}}k(\texttt{inputs})\cdot q(\texttt{slots})^T, \texttt{axis='inputs'}\right)$$

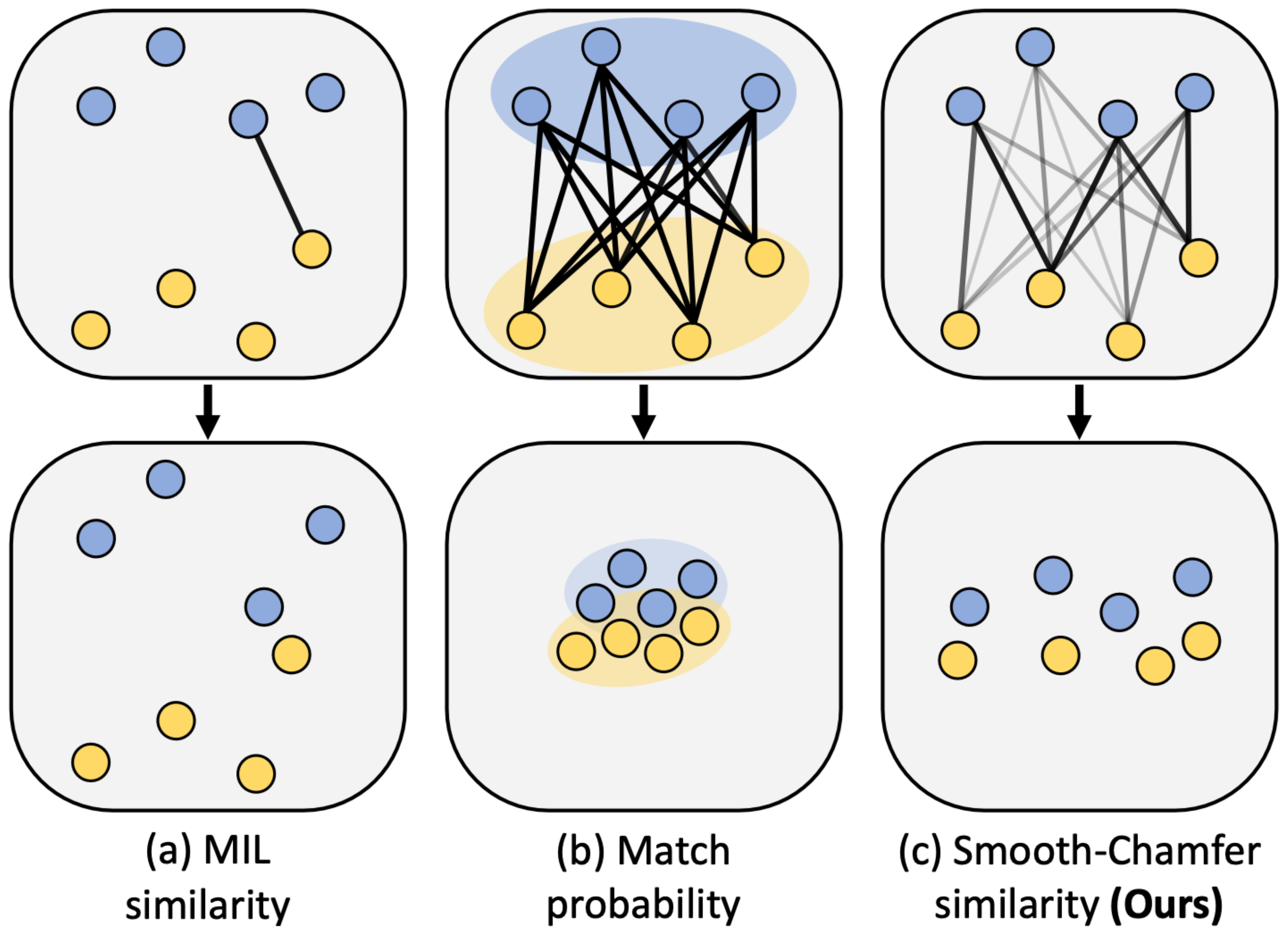
Conventional transformer attention scheme

- This way of normalization lets slots ***compete*** with each other.
- Each slot attends to nearly disjoint sets of local features, and these sets will correspond to the ***distinctive semantics***.

Overall architecture

3. Smooth-Chamfer similarity



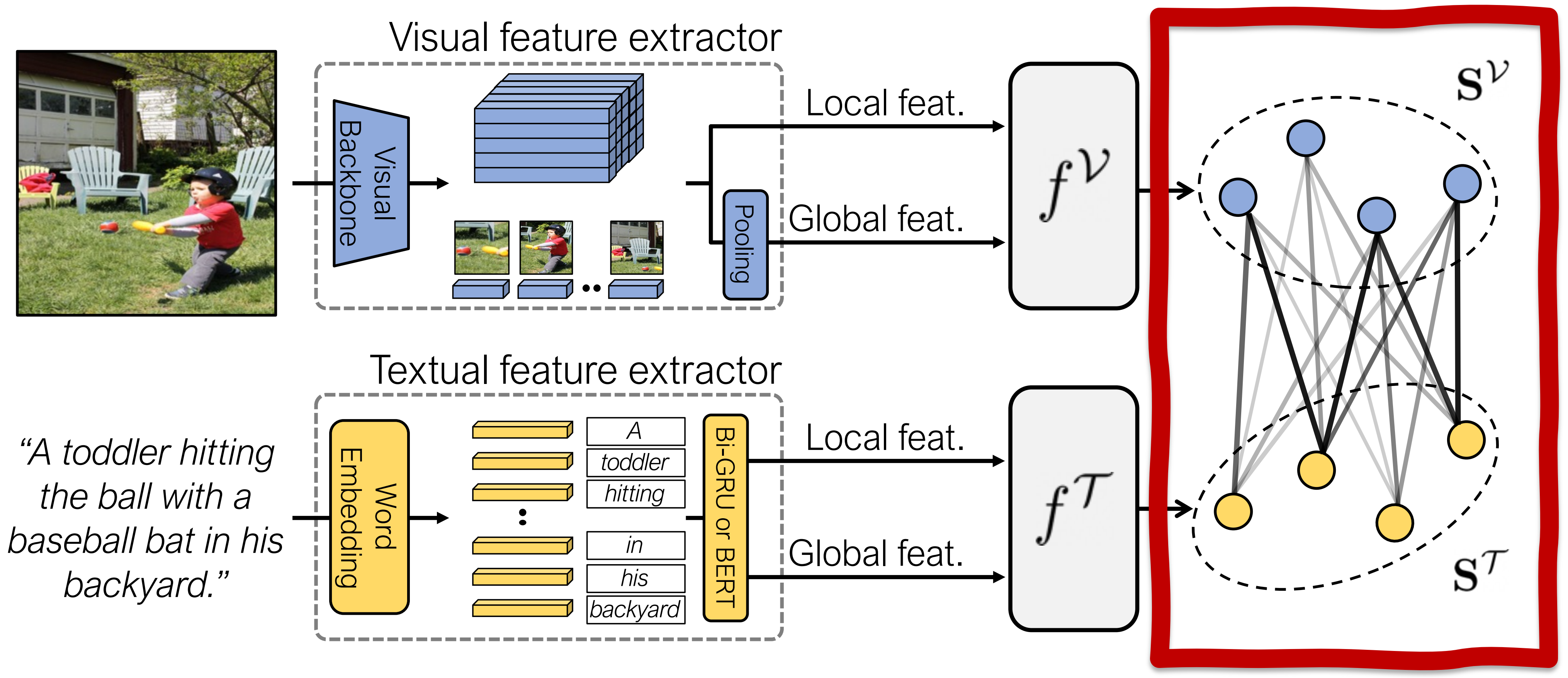


$$s_{\mathrm{SC}}\left(\mathbf{S}_1, \mathbf{S}_2\right) = \frac{1}{2\alpha\left|\mathbf{S}_1\right|} \sum_{x \in \mathbf{S}_1} \operatorname*{LSE}_{y \in \mathbf{S}_2}(\alpha c(x, y)) + \frac{1}{2\alpha\left|\mathbf{S}_2\right|} \sum_{y \in \mathbf{S}_2} \operatorname*{LSE}_{x \in \mathbf{S}_1}(\alpha c(x, y))$$

- SC similarity associates every possible pair (→***dense supervision***) with different degree of weights (→***no set collapsing***)

Overall architecture

3. Smooth-Chamfer similarity



$$s_{\mathrm{SC}}\left(\mathbf{S}_1, \mathbf{S}_2\right) = \frac{1}{2\alpha\left|\mathbf{S}_1\right|} \sum_{x \in \mathbf{S}_1} \underset{y \in \mathbf{S}_2}{\mathrm{LSE}}(\alpha c(x, y)) + \frac{1}{2\alpha\left|\mathbf{S}_2\right|} \sum_{y \in \mathbf{S}_2} \underset{x \in \mathbf{S}_1}{\mathrm{LSE}}(\alpha c(x, y))$$

Gradient respect to similarity between elements

$$\frac{\partial s_{\mathrm{SC}}\left(\mathbf{S}_1, \mathbf{S}_2\right)}{\partial c\left(x', y'\right)} = \frac{1}{2\left|\mathbf{S}_1\right|} \frac{e^{\alpha c\left(x', y'\right)}}{\sum_{y \in \mathbf{S}_2} e^{\alpha c\left(x', y\right)}} + \frac{1}{2\left|\mathbf{S}_2\right|} \frac{e^{\alpha c\left(x', y'\right)}}{\sum_{x \in \mathbf{S}_1} e^{\alpha c\left(x, y'\right)}}$$

- Gradients for the elements pair are determined by the ***relative proximity***.
- This weighting scheme enables ***dense supervision without collapsing***.

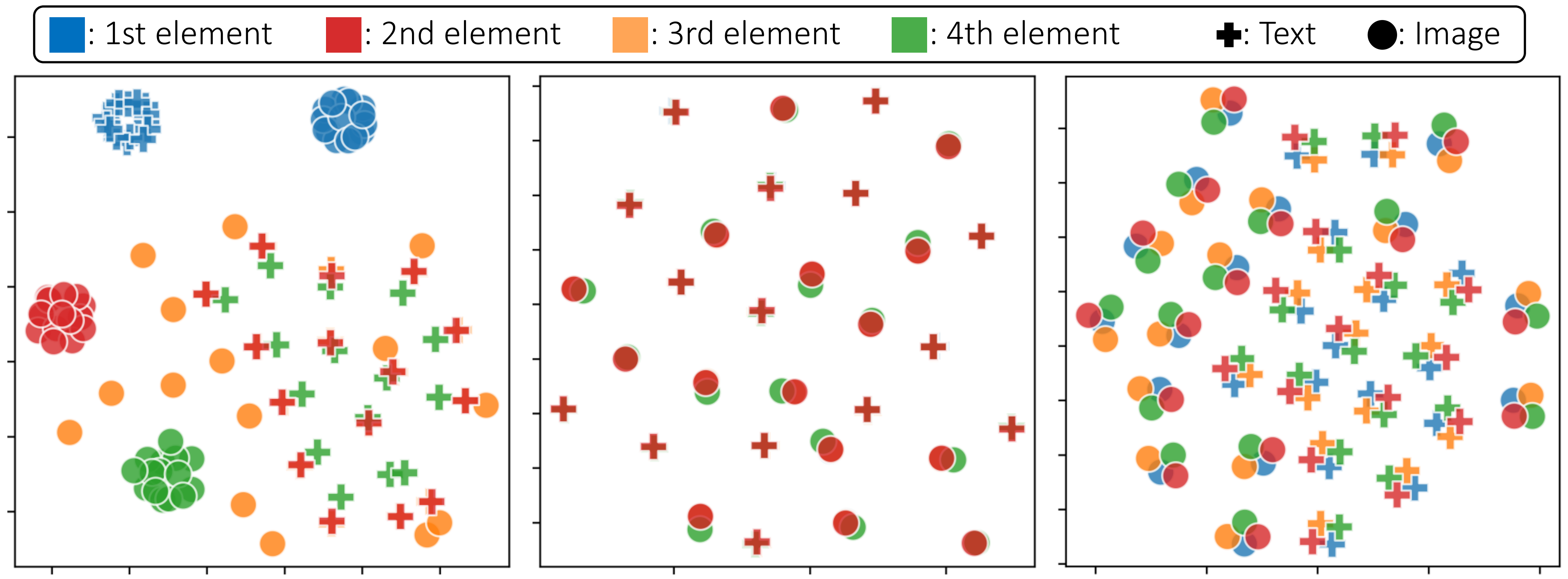
Experiments

		1K Test Images							5K Test Images						
		Image-to-Text			Text-to-Image				Image-to-Text			Text-to-Image			
Method	CA	R@1	R@5	R@10	R@1	R@5	R@10	RSUM	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
ResNet-152 + Bi-GRU															
VSE++ [17]	✗	64.6	90.0	95.7	52.0	84.3	92.0	478.6	41.3	71.1	81.2	30.3	59.4	72.4	355.7
PVSE [45]	✗	69.2	91.6	96.6	55.2	86.5	93.7	492.8	45.2	74.3	84.5	32.4	63.0	75.0	374.4
PCME [10]	✗	68.8	-	-	54.6	-	-	-	44.2	-	-	31.9	-	-	-
Ours	✗	70.3	91.5	96.3	56.0	85.8	93.3	**493.2**	47.2	74.8	84.1	33.8	63.1	74.7	**377.7**
Faster R-CNN + Bi-GRU															
SCAN† [30]	✓	72.7	94.8	98.4	58.8	88.4	94.8	507.9	50.4	82.2	90.0	38.6	69.3	80.4	410.9
VSRN† [31]	✗	76.2	94.8	98.2	62.8	89.7	95.1	516.8	53.0	81.1	89.4	40.5	70.6	81.1	415.7
CAAN [53]	✓	75.5	95.4	98.5	61.3	89.7	95.2	515.6	52.5	83.3	90.9	41.2	70.3	82.9	421.1
IMRAM† [6]	✓	76.7	95.6	98.5	61.7	89.1	95.0	516.6	53.7	83.2	91.0	39.7	69.1	79.8	416.5
SGRAF† [14]	✓	79.6	96.2	98.5	63.2	90.7	96.1	524.3	57.8	-	91.6	41.9	-	81.3	-
VSE$_\infty$ [27]	✗	78.5	96.0	98.7	61.7	90.3	95.6	520.8	56.6	83.6	91.4	39.3	69.9	81.1	421.9
NAAF† [52]	✓	80.5	96.5	98.8	64.1	90.7	96.5	527.2	58.9	85.2	92.0	42.5	70.9	81.4	430.9
Ours	✗	79.8	96.2	98.6	63.6	90.7	95.7	524.6	58.8	84.9	91.5	41.1	72.0	82.4	430.7
Ours†	✗	80.6	96.3	98.8	64.7	91.4	96.2	**528.0**	60.4	86.2	92.4	42.6	73.1	83.1	**437.8**
ResNeXt-101 + BERT															
VSE$_\infty$ [27]	✗	84.5	98.1	99.4	72.0	93.9	97.5	545.4	66.4	89.3	94.6	51.6	79.3	87.6	468.9
VSE$_\infty^\dagger$ [27]	✗	85.6	98.0	99.4	73.1	94.3	97.7	548.1	68.1	90.2	95.2	52.7	80.2	88.3	474.8
Ours	✗	86.3	97.8	99.4	72.4	94.0	97.6	547.5	69.1	90.7	95.6	52.1	79.6	87.8	474.9
Ours†	✗	86.6	98.2	99.4	73.4	94.5	97.8	**549.9**	71.0	91.8	96.3	53.4	80.9	88.6	**482.0**

		Image-to-text			Text-to-image			
Method	CA	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
ResNet-152 + Bi-GRU								
VSE++	✗	52.9	80.5	87.2	39.6	70.1	79.5	409.8
PVSE*	✗	59.1	84.5	91.0	43.4	73.1	81.5	432.6
PCME*	✗	58.5	81.4	89.3	44.3	72.7	81.9	428.1
Ours	✗	61.8	85.5	91.1	46.1	74.8	83.3	**442.6**
Faster R-CNN + Bi-GRU								
SCAN†	✓	67.4	90.3	95.8	48.6	77.7	85.2	465.0
VSRN†	✗	71.3	90.6	96.0	54.7	81.8	88.2	482.6
CAAN	✓	70.1	91.6	97.2	52.8	79.0	87.9	478.6
IMRAM†	✓	74.1	93.0	96.6	53.9	79.4	87.2	484.2
SGRAF†	✓	77.8	94.1	97.4	58.5	83.0	88.8	499.6
VSE$_\infty$	✗	76.5	94.2	97.7	56.4	83.4	89.9	498.1
NAAF†	✓	81.9	96.1	98.3	61.0	85.3	90.6	**513.2**
Ours	✗	77.8	94.0	97.5	57.5	84.0	90.0	500.8
Ours†	✗	80.9	94.7	97.6	59.4	85.6	91.1	509.3
ResNeXt-101 + BERT								
VSE$_\infty$	✗	88.4	98.3	99.5	74.2	93.7	96.8	550.9
VSE$_\infty^\dagger$	✗	88.7	98.9	99.8	76.1	94.5	97.1	555.1
Ours	✗	88.8	98.5	99.6	74.3	94.0	96.7	551.9
Ours†	✗	90.6	99.0	99.6	75.9	94.7	97.3	**557.1**

- Achieves the state-of-the-art on various benchmarks and settings
- Outperforms some of the previous work that requires x80 FLOPs

Experiments



(a) MIL (b) Match Probability (c) Ours

Thank you!

Poster: THU-PM-269