Linguistic-Aware Patch Slimming Framework for Fine-grained Cross-Modal Alignment

Zheren Fu · Lei Zhang · Hou Xia · Zhendong Mao

University of Science and Technology of China, Hefei, China

Background

- ✓ **Cross-modal alignment** aims to bridge the semantic gap between different modalities, such as visual and linguistic ones. It is a fundamental technology for many multi-modal tasks, including image-text retrieval, visual question answering, image captioning.

- ✓ The critical challenge of cross-modal alignment lies in efficiently **measuring the semantic similarities** between images and texts to achieve a high-quality alignment.

Background

- ✓ In general, existing cross-modal alignments can be classified into **two paradigms**.
- ✓ The first **coarse-grained alignment** separately encodes the whole images and texts into a unified embedding space, then directly computes the similarity of global embeddings..
- ✓ The second **fine-grained alignment** applies cross-modal interaction between visual and textual local features, then aggregates local alignments to learn a cumulative similarity.

coarse-grained

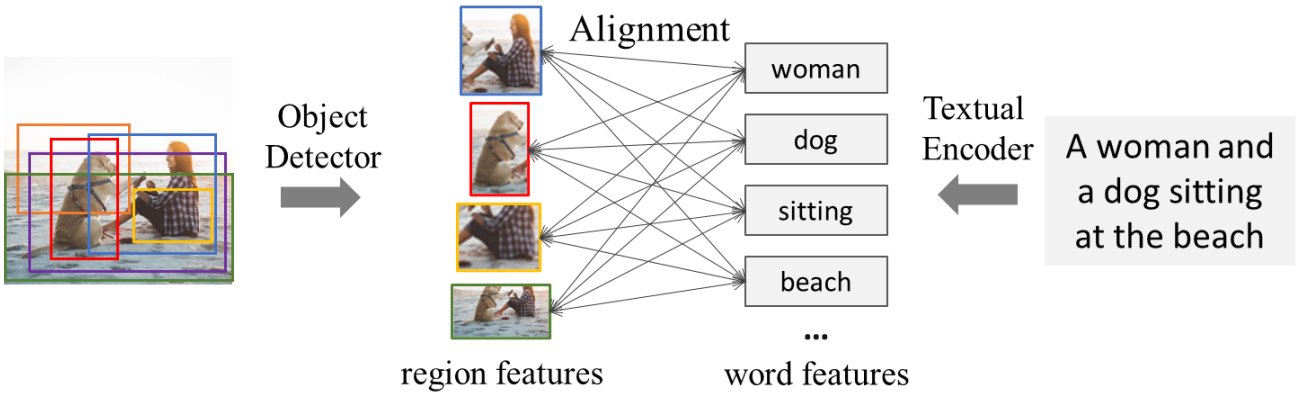
fine-grained

Motivation

✓ Previous fine-grained methods adhered to the detector-based roadmap, which entails a **two-step process**:

(1) Extract region features through a pre-trained object detector, e.g., Faster-RCNN.

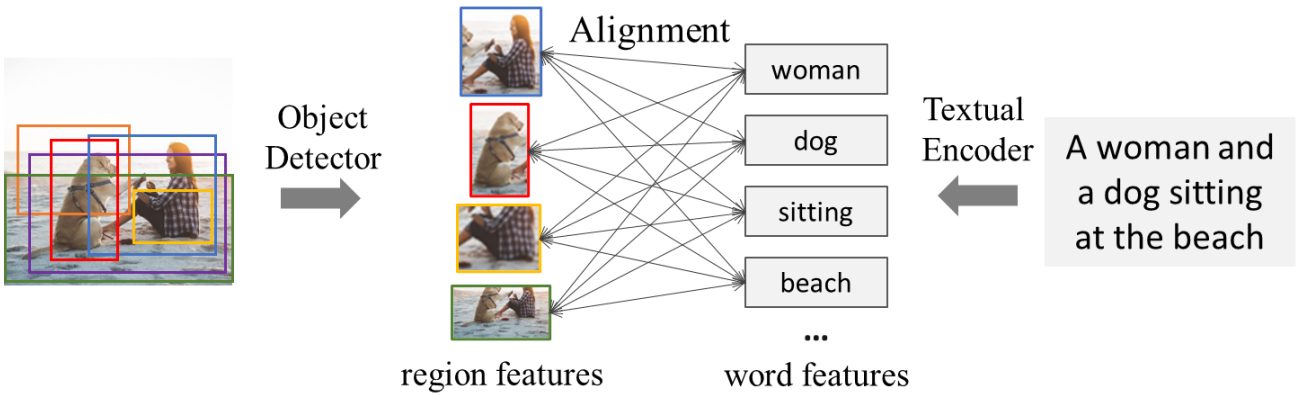
(2) Compute region-word alignments between visual regions and textual words.



Previous Detector-based region-word alignment

Motivation

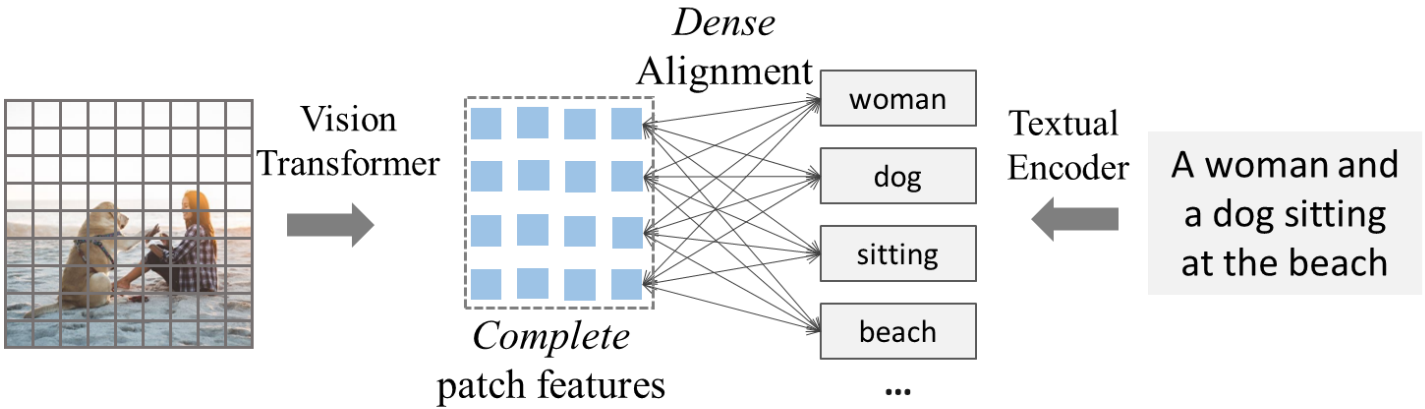
- ✓ The detector is **compute-expensive** by complex region detection during inference (e.g., RPNs, RoI Align, NMS operations)
- ✓ Besides, detectors usually cannot participate in the end-to-end training and **bring error propagation issues**.



Previous Detector-based region-word alignment

Motivation

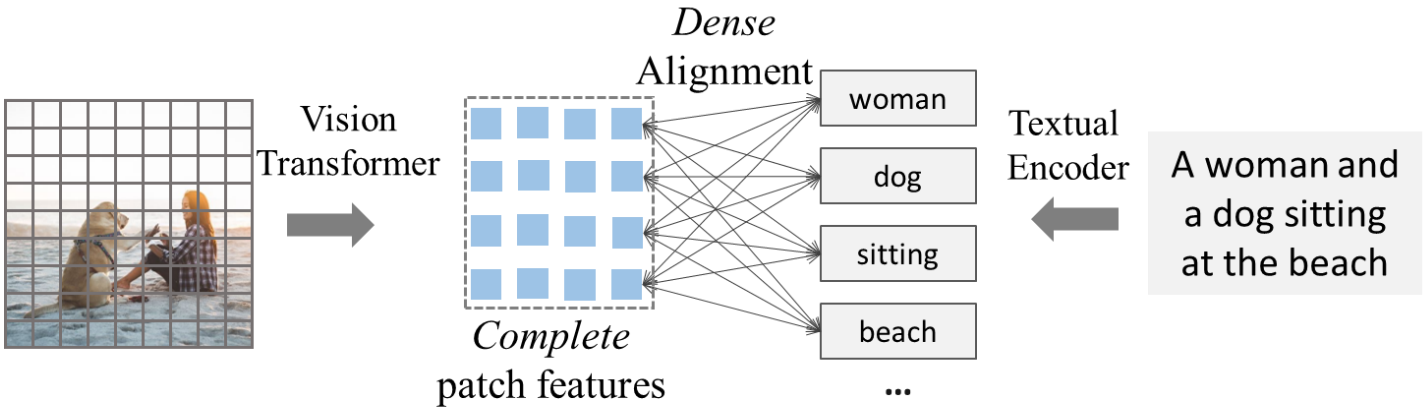
- ✓ Recent works adopt **pure transformer architectures**, e.g., vision transformer, to divide images into non-overlapping patches and encode patch features to construct patch-word alignments.
- ✓ The transformer-based method is a flexible **end-to-end training** framework, efficient for feature extraction, owns scalable performances compared to the detector-based, and has become mainstream.



Vanilla Transformer-based patch-word alignment

Motivation

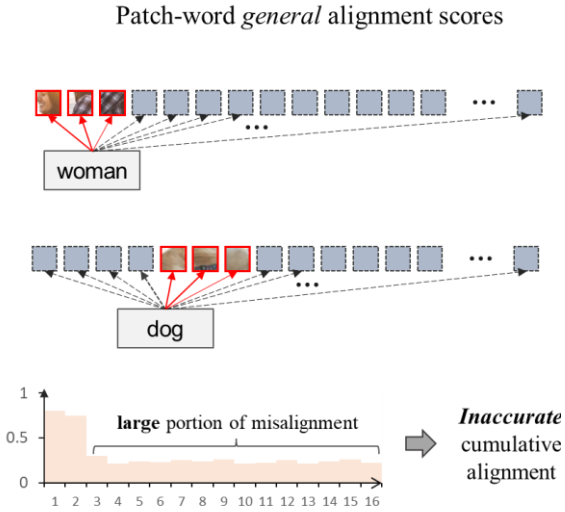
- ✓ However, the vanilla transformer-based patch-word framework has inherent defects, **the patch redundancy and patch ambiguity** problems for semantic alignment.



Vanilla Transformer-based patch-word alignment

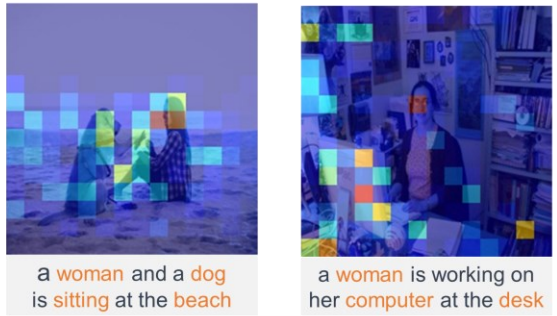
Motivation: patch redundancy

- ✓ The vision transformer divides images into minute patches (at 224 and 284 image resolutions, it generates 14×14=196 and 24×24=576 patches), a substantial proportion of them proves to be redundant, e.g., non-salient backgrounds or text-irrelevant areas.



(b) Patch redundancy

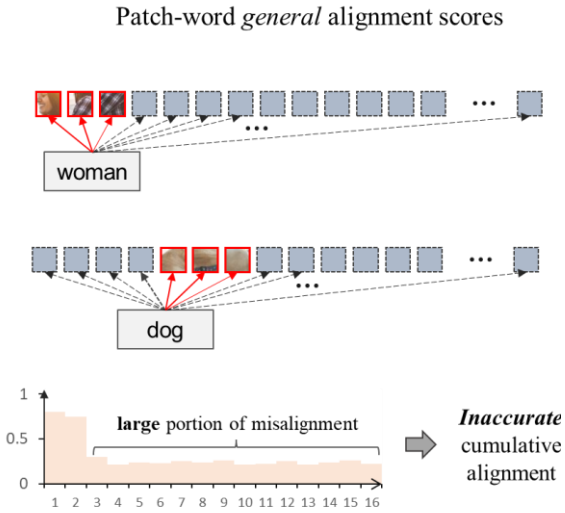
(a) Patch Attention with image itself



(b) Patch Attention with image-text alignment

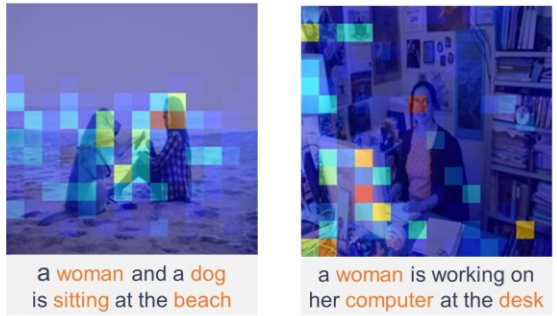
Motivation: patch redundancy

- ✓ The massive redundant patches will overshadow crucial visual patches, and accumulate unbearable misalignment during the patch-word interaction, ultimately bringing inaccurate cumulative alignments.



(b) Patch redundancy

(a) Patch Attention with image itself

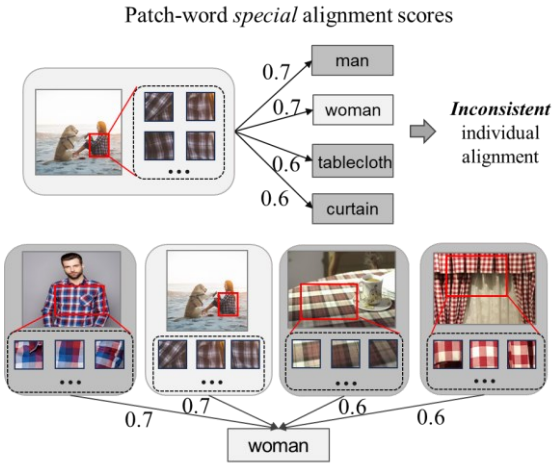


(b) Patch Attention with image-text alignment

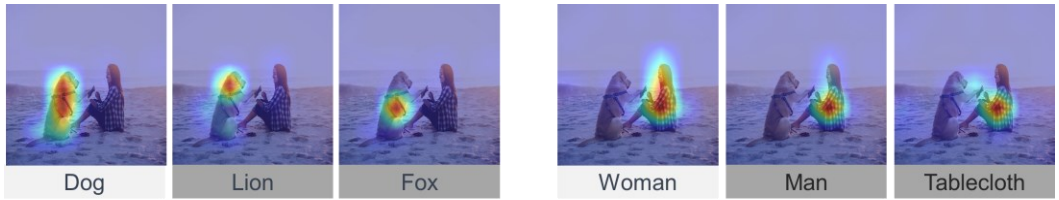
Motivation: patch ambiguity

- ✓ More importantly, these fragmented patches are tiny components of an image. Tiny patches lack semantic integrity compared to complete visual regions. It will lead to ambiguous semantic expressions.

- ✓ Visual patches always get moderate alignment scores for distinct language concepts, which brings inconsistent patch-word alignment in local regions with negative image-text pairs.



(c) Patch ambiguity

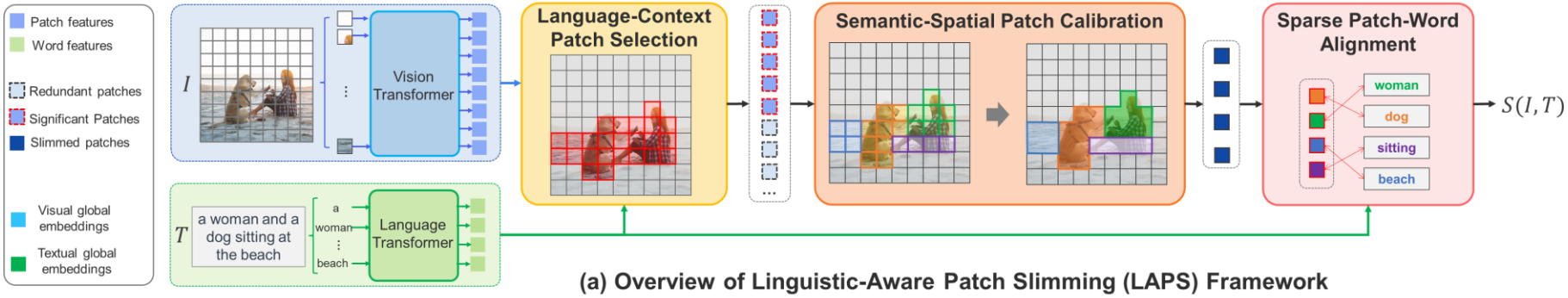


(c) Patch Attention with different linguistic words

Contribution

- ✓ Consequently, how to ensure the semantic integrity of visual patches to establish accurate alignment, is a core issue for transformer-based cross-modal frameworks.

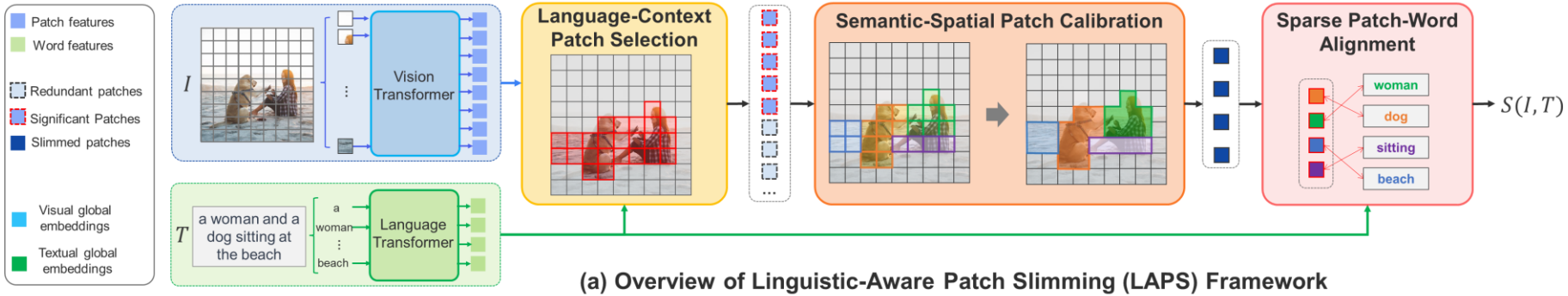
- ✓ To address these problems, we introduce a **Linguistic-Aware Patch Slimming** (LAPS) framework, which effectively eliminates extensive redundant patches through linguistic supervision, and calibrates the semantic and structural information for significant patches to transform an average semantic expression into an optimal semantic for a certain image.



(a) Overview of Linguistic-Aware Patch Slimming (LAPS) Framework

Contribution

- ✓ To the best of our knowledge, LAPS is the first explicitly explore visual **patch selection and patch calibration** with language contexts to facilitate patch-word alignment.
- ✓ Therefore, LAPS extends the vanilla transformer-based framework **to achieve more accurate and consistent** patch-word alignments.



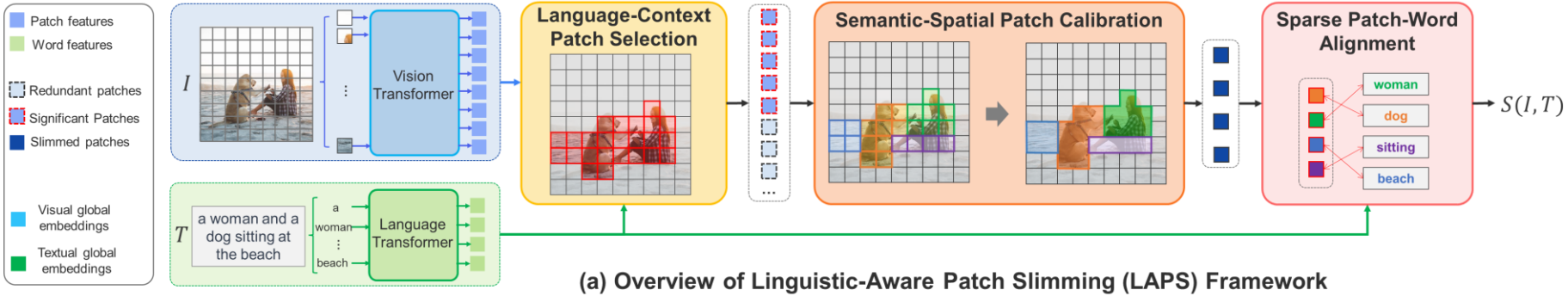
(a) Overview of Linguistic-Aware Patch Slimming (LAPS) Framework

Framework

(1) We first effectively estimate the semantic significance of visual patches using the Language-Context Patch Selection (LPS) module to pick out significant patches with differentiable sampling.

(2) Next, we adaptively rectify the semantic and structural information for significant patches through the Semantic-Spatial Patch Calibration (SPC) module to obtain distinct semantic expressions.

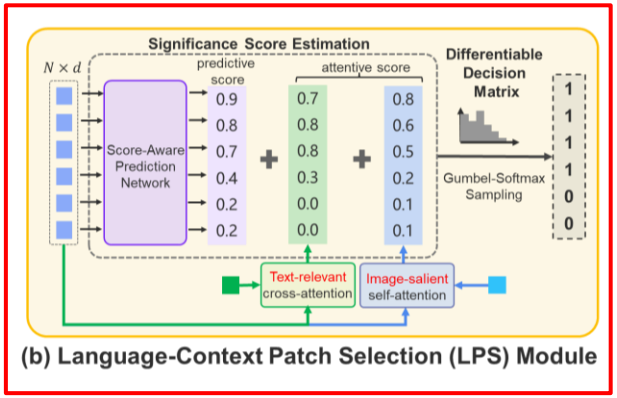
(3) Finally, we employ the Sparse Patch-Word Alignment (SPA) module to facilitate the fine-grained interaction between visual patches and textual words.



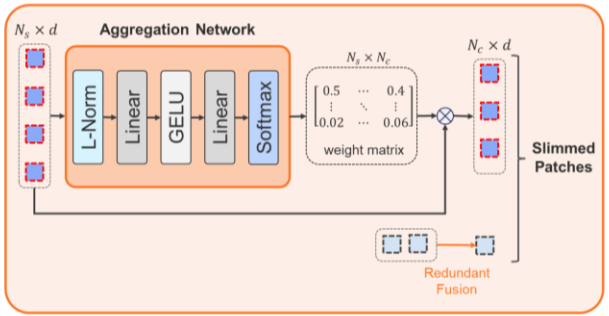
(a) Overview of Linguistic-Aware Patch Slimming (LAPS) Framework

Language-Context Patch Selection

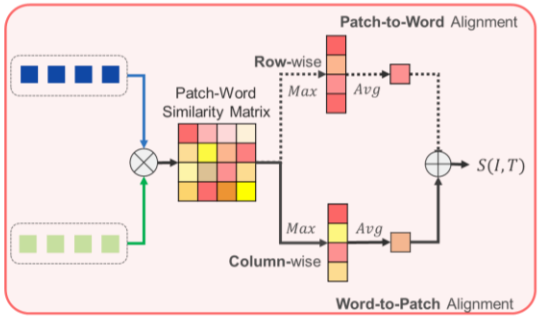
- ✓ We estimate the patch significance scores to identify redundant visual patches through linguistic supervision and select significant ones using differentiable sampling.



(b) Language-Context Patch Selection (LPS) Module



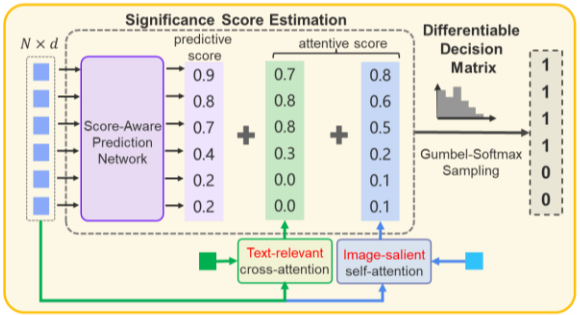
(c) Semantic-Spatial Patch Calibration (SPC) Module



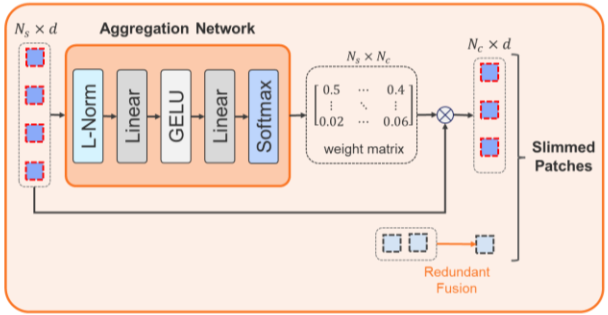
(d) Sparse Patch-Word Alignment (SPA) Module

Semantic-Spatial Patch Calibration

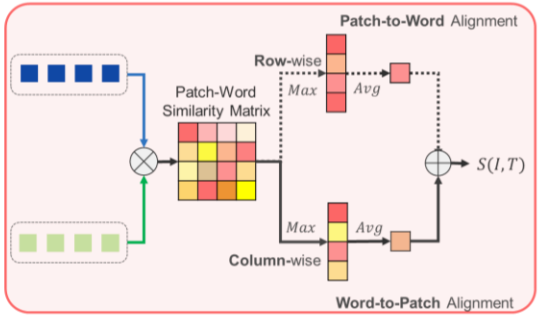
- ✓ We rectify significant visual patches with semantic and spatial relationships to obtain semantic integrity and structural information comparable to linguistic texts.



(b) Language-Context Patch Selection (LPS) Module



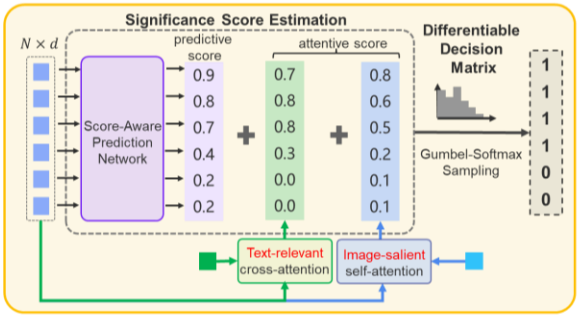
(c) Semantic-Spatial Patch Calibration (SPC) Module



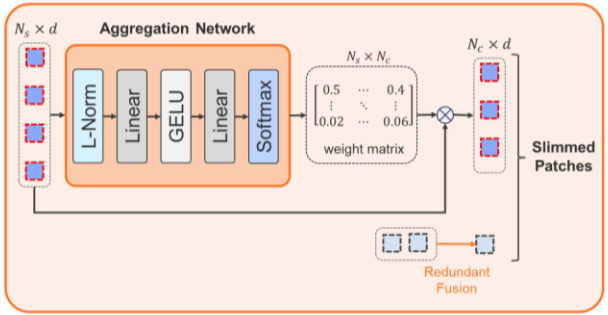
(d) Sparse Patch-Word Alignment (SPA) Module

Sparse Patch-Word Alignment

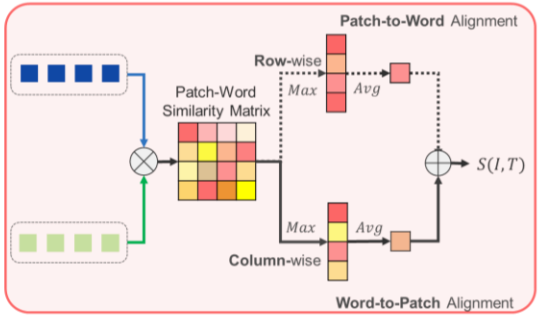
- ✓ Finally, we employ the Sparse Patch-Word Alignment (SPA) module to facilitate the fine-grained interaction between visual patches and textual words.



(b) Language-Context Patch Selection (LPS) Module



(c) Semantic-Spatial Patch Calibration (SPC) Module



(d) Sparse Patch-Word Alignment (SPA) Module

Experiments

- ✓ Extensive experiments on Flickr30K and MS-COCO show our proposed method outperforms the state-of-the-art methods.

Table 1. Comparisons of image-text retrieval performances on Flickr30K and MS-COCO test-set. We list the details of feature encoders, image resolution, and the number of obtained regions/patches by visual encoders (e.g., 'ViT-Base-224' represents the base–version of Vision Transformer [8] with 224×224 image resolution input, regarding 16×16 pixels as one patch, and getting 14×14 visual patches for one image). *FG* indicates whether it is the fine-grained cross-modal alignment. The best results are marked **bold**.

Method	FG	Flickr30K 1K							MS-COCO 1K							MS-COCO 5K						
		Image-to-Text			Text-to-Image			rSum	Image-to-Text			Text-to-Image			rSum	Image-to-Text			Text-to-Image			rSum
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
***Faster R-CNN + BERT–Base*, 36 pre-computed regions**																						
HREM [14]	✗	83.3	96.0	98.1	63.5	87.1	92.4	520.4	81.1	96.6	98.9	66.1	91.6	96.5	530.7	62.3	87.6	93.4	43.9	73.6	83.3	444.1
TGDT [30]	✓	61.3	86.0	91.4	76.8	93.2	96.4	505.1	65.4	91.8	96.5	78.5	96.4	98.9	527.5	43.3	73.5	83.3	57.5	84.8	91.6	434.0
CHAN [35]	✓	80.6	96.1	97.8	63.9	87.5	92.6	518.5	81.4	96.9	98.9	66.5	92.1	96.7	532.6	59.8	87.2	93.3	44.9	74.5	84.2	443.9
***ViT-Base-224 + BERT–base*, 14×14 patches**																						
VSE++ [9]	✗	71.8	92.8	96.5	59.4	84.7	90.9	496.1	75.0	94.6	98.0	62.7	89.4	94.9	514.6	52.4	80.3	88.8	40.6	70.4	81.1	413.4
SCAN [21]	✓	69.5	90.9	95.6	56.4	83.1	90.0	485.6	76.0	95.4	98.1	64.5	90.8	95.8	520.6	53.9	81.8	90.0	42.9	72.3	82.5	423.5
SGR [7]	✓	69.7	90.8	95.2	59.1	84.1	89.9	488.7	77.2	95.0	98.0	65.1	90.7	95.8	521.8	54.9	82.8	90.5	42.8	72.2	82.5	425.8
CHAN [35]	✓	69.2	91.8	95.0	58.4	84.9	90.6	489.9	77.1	95.1	98.1	65.0	91.0	96.0	522.2	56.3	83.2	90.1	43.0	72.6	82.8	428.0
LAPS	✓	**74.0**	**93.4**	**97.4**	**62.5**	**87.3**	**92.7**	**507.3**	**78.7**	**95.5**	**98.3**	**66.2**	**91.3**	**96.2**	**526.3**	**57.5**	**84.0**	**90.8**	**44.5**	**74.0**	**83.6**	**434.4**
***ViT-Base-384 + BERT–base*, 24×24 patches**																						
VSE++ [9]	✗	77.1	95.7	97.5	65.8	90.2	94.3	520.5	77.0	95.7	98.4	64.6	91.1	96.2	523.0	54.9	82.8	90.4	42.4	72.4	82.8	425.8
SCAN [21]	✓	75.4	94.4	96.9	63.6	88.6	93.5	512.5	76.1	95.5	98.5	65.1	91.6	96.3	523.1	53.3	81.8	90.0	42.6	72.6	82.9	423.1
SGR [7]	✓	76.9	94.9	98.1	64.2	88.4	93.3	515.8	75.8	95.7	98.6	65.6	92.0	96.5	524.2	53.3	81.0	89.6	42.9	73.1	83.7	423.6
CHAN [35]	✓	75.4	94.5	97.6	63.2	88.6	93.1	512.4	78.1	95.8	98.6	66.1	92.1	96.6	527.3	55.6	83.8	91.2	43.4	73.6	83.5	431.1
LAPS	✓	**79.0**	**96.0**	**98.1**	**67.3**	**90.5**	**94.5**	**525.4**	**78.6**	**96.3**	**98.9**	**68.0**	**92.4**	**96.8**	**531.0**	**57.4**	**84.9**	**92.5**	**46.4**	**75.8**	**85.2**	**442.2**
***Swin-Base-224 + BERT–base*, 7×7 patches**																						
VSE++ [9]	✗	82.5	96.5	98.9	70.0	91.4	95.1	534.4	83.3	97.5	99.3	71.0	93.0	96.7	540.9	64.0	88.2	94.2	49.9	78.0	86.6	460.9
SCAN [21]	✓	79.0	95.9	98.2	67.7	90.6	94.9	526.3	80.9	97.0	99.1	69.7	93.1	97.1	536.9	60.7	86.6	93.2	48.1	77.1	86.1	451.8
SGR [7]	✓	80.4	97.0	98.7	66.9	90.2	94.5	527.6	81.2	97.1	99.1	69.9	93.2	97.2	537.7	61.0	86.7	93.2	48.6	77.2	86.3	453.1
CHAN [35]	✓	81.4	97.0	98.6	68.5	90.6	94.5	530.6	81.6	97.2	99.3	70.6	93.7	97.6	539.8	64.1	87.9	93.5	49.1	77.3	86.1	458.0
LAPS	✓	**82.4**	**97.4**	**99.5**	**70.0**	**91.7**	**95.4**	**536.3**	**84.0**	**97.6**	**99.3**	**72.1**	**93.7**	**97.3**	**544.1**	**64.5**	**89.2**	**94.4**	**51.6**	**78.9**	**87.2**	**465.8**
***Swin-Base-384 + BERT–base*, 12×12 patches**																						
VSE++ [9]	✗	83.3	97.5	99.2	71.1	93.2	96.2	540.6	82.9	97.7	99.4	71.3	93.5	97.3	542.1	63.0	88.5	94.3	50.1	78.9	87.4	462.2
SCAN [21]	✓	81.9	96.9	98.9	70.0	92.7	95.8	536.1	81.6	96.8	99.1	69.1	92.7	96.7	536.1	61.1	87.3	93.3	47.8	76.9	85.9	452.4
SGR [7]	✓	80.7	96.8	99.0	69.9	91.7	95.3	533.4	81.9	96.7	99.1	69.3	92.8	96.7	536.6	62.8	87.0	92.9	48.1	77.0	86.0	453.8
CHAN [35]	✓	81.2	96.7	98.8	70.3	92.2	95.9	535.0	83.1	97.3	99.2	70.4	93.1	97.1	540.2	63.4	88.4	94.1	49.2	77.9	86.6	459.5
LAPS	✓	**85.1**	**97.7**	**99.2**	**74.0**	**93.0**	**96.3**	**545.3**	**84.1**	**97.4**	**99.2**	**72.1**	**93.9**	**97.4**	**544.1**	**67.1**	**88.6**	**94.3**	**53.0**	**79.5**	**87.6**	**470.1**

Experiments

- ✓ Extensive experiments on Flickr30K and MS-COCO show our proposed method outperforms the state-of-the-art methods.

Table 2. The comparisons of image-text retrieval for Vision-Language Pre-training (VLP) Models. *FG* indicates whether it is the fine-grained alignment. # represents the zero-shot learning.

Method	*FG*	Flickr30K 1K				MS-COCO 5K			
		Image-to-Text		Text-to-Image		Image-to-Text		Text-to-Image	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
UNITER [5]	✓	87.3	98.0	75.6	94.1	65.7	88.6	52.9	79.9
VILT [19]	✓	83.5	96.7	64.4	88.7	61.5	86.3	42.7	72.9
SOHO [16]	✓	86.5	98.1	72.5	92.7	66.4	88.2	50.6	78.0
ALBEF [24]	✓	95.9	99.8	85.6	97.5	77.6	94.3	60.7	84.3
BLIP [25]	✓	96.6	99.8	87.2	97.5	80.6	95.2	63.1	85.3
***CLIP-ViT-Base-224 + CLIP-BERT–Base*, 14×14 patches**									
CLIP$^{\#}$ [36]	✗	81.4	96.2	61.1	85.4	52.3	76.2	33.3	58.2
VSE++ [9]	✗	92.2	99.1	80.5	**95.6**	66.8	88.2	53.6	79.7
SCAN [21]	✓	88.2	98.1	75.3	93.1	65.4	88.0	50.7	77.6
LAPS	✓	**92.9**	**99.3**	**80.6**	**95.5**	**69.8**	**90.4**	**54.3**	**80.0**
***CLIP-ViT-Large-224 + CLIP-BERT–Large*, 16×16 patches**									
CLIP$^{\#}$ [36]	✗	85.0	97.7	64.3	87.0	55.9	79.1	35.9	60.9
VSE++ [9]	✗	94.0	99.5	83.4	96.4	68.5	89.4	56.7	81.9
SCAN [21]	✓	90.0	98.5	81.0	95.9	68.0	90.4	53.2	80.7
LAPS	✓	**94.6**	**99.9**	**84.9**	**97.3**	**72.9**	**91.7**	**57.1**	**81.3**

Ablation Study

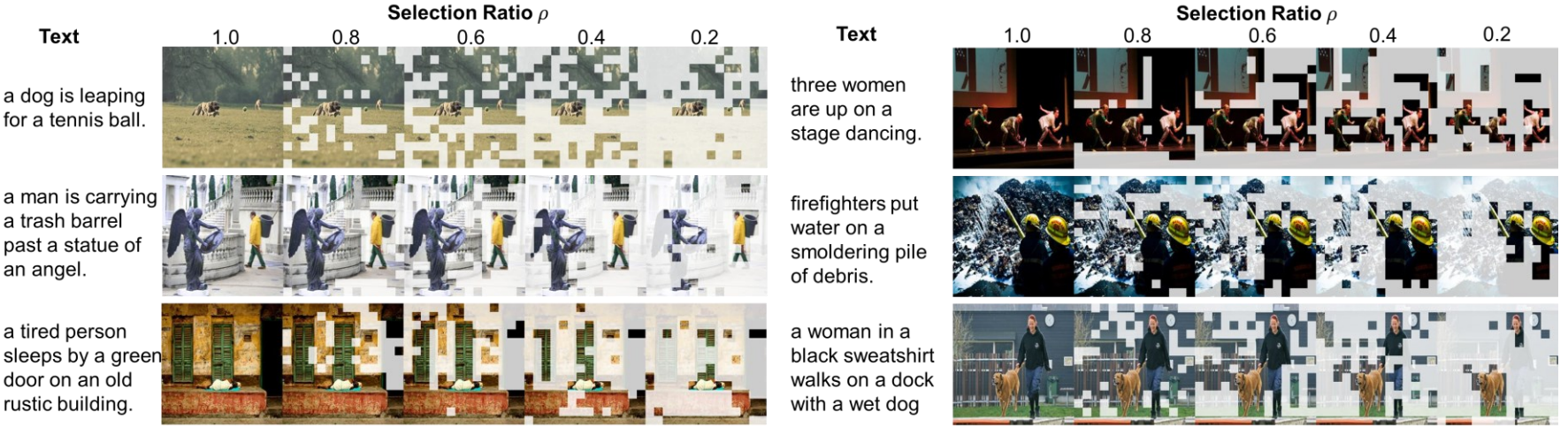
✓ The ablation study show our patch slimming methods are meaningful.

Table 4. Comparison of different module ablations for our framework on Flickr30K. We also show the results of the word slimming (selection + aggregation) of textual modality for our framework.

Modules	Different Settings	IMG → TEXT		TEXT → IMG	
		R@1	R@5	R@1	R@5
LPS	without patch selection process	69.2	91.9	58.5	84.9
	without language-context	71.1	92.2	59.4	85.5
	only attentive scores	73.5	93.1	61.9	86.8
SPC	without patch calibration process	70.4	91.3	58.9	85.3
	without redundant fusion	73.5	93.2	61.1	87.2
	use the clustering algorithm [46]	68.4	88.5	57.0	82.6
SPA	replace with SCAN alignment [21]	71.3	91.4	60.8	85.6
	only patch-to-word alignment	70.9	90.8	58.9	85.1
	only word-to-patch alignment	72.7	92.5	60.3	86.4
	introduce word selection	70.1	90.3	57.5	82.7
	introduce word aggregation	71.3	91.6	58.8	84.3
	introduce word selection & aggregation	67.7	88.2	55.1	80.5
	Complete LAPS	**74.0**	**93.4**	**62.5**	**87.3**

Visualization

- ✓ The visualization of selected patches with associated texts under different selection ratios ρ on Flickr30K. Our framework can gradually focus on more salient and text-relevant areas in images as selection ratios decrease and have better interpretability.

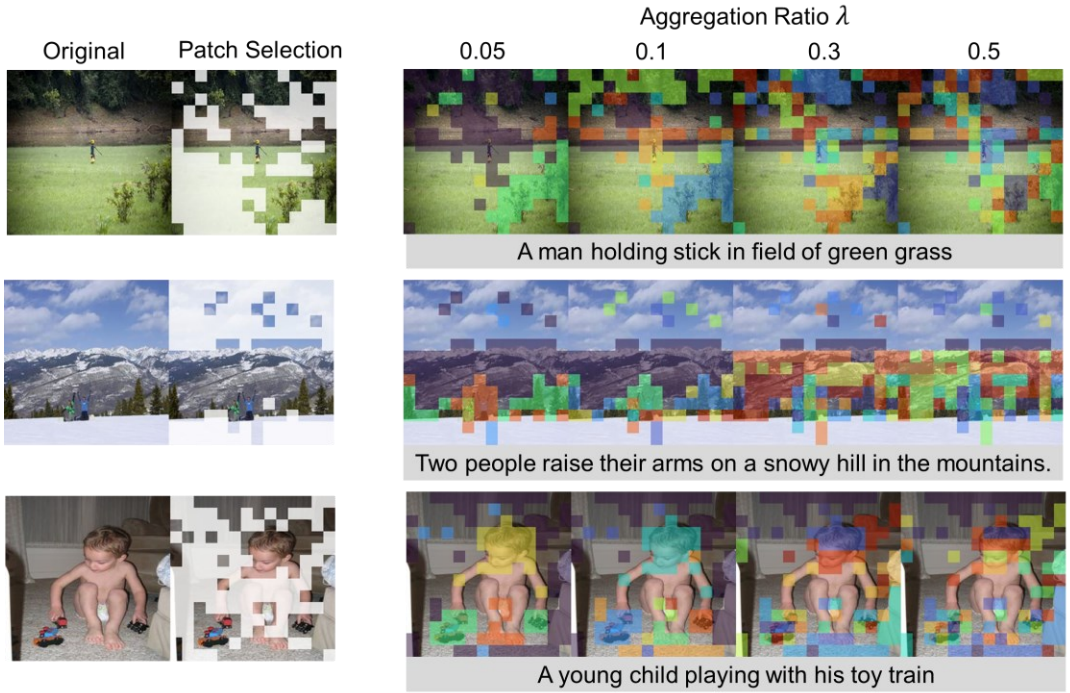


Visualization

- ✓ The visualization of selected patches with the different language contexts and supervision. The texts below the images describe the semantic contents of images with various perspectives.

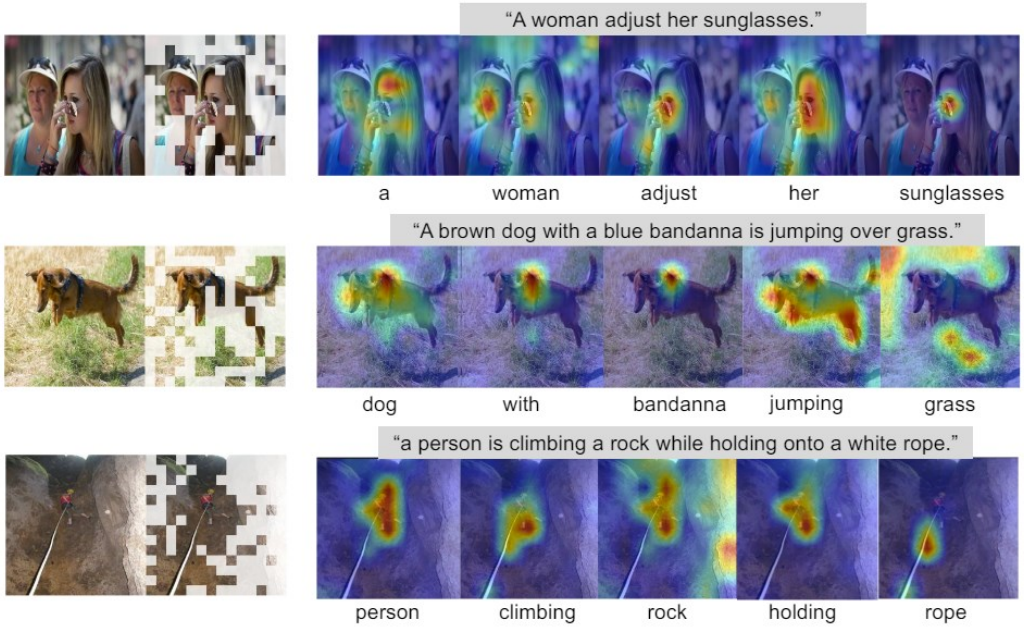
Visualization

- ✓ The visualization of aggregated patches with the different aggregation ratios λ. The patches are merged into complete regions with clear semantics and have better interpretability.



Visualization

- ✓ The visualization of fine-grained patch-word alignment with each linguistic word. We show the alignment maps by the gradient-weighted attention on original images.



Conclusion

- ✓ In this paper, we introduce a novel Linguistic-Aware Patch Slimming framework (LAPS) for cross-modal alignment, which is the first work that explicitly focuses on patch-word alignment on pure transformer-based architectures to solve patch redundancy and ambiguity problems.
- ✓ LAPS identifies significant visual patches with language supervision and then rectifies the semantic and structural information to construct more accurate and consistent alignment.
- ✓ Extensive experiments on various benchmarks and visual encoders demonstrate the superiority of our framework.