Image Difference Captioning with Pre-training and Contrastive Learning

Linli Yao, Weiying Wang, Qin Jin

AIM3 Lab, School of Information, Renmin University of China

AI · M3 中国人民大学多媒体计算实验室

中国人民大學 RENMIN UNIVERSITY OF CHINA

Outline

- Task Introduction
- Method
- Experiments and Analysis
- Conclusions

Image Difference Captioning

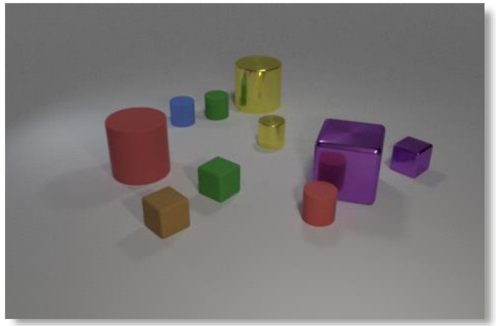
Image Difference Captioning(IDC) task aims to *describe the visual differences* between two similar images *with natural language*.

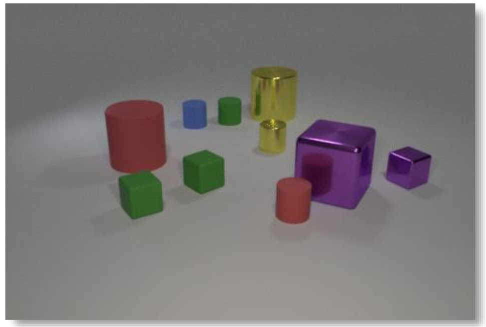
" Animal1 is covered in **yellow** , **green** and **orange** feathers , while animal2 is covered in **greenish grey** feathers with **dark orange** feathers on abdomen and chest ."

Assist ornithologists to distinguish similar species, report salient changes in surveillance

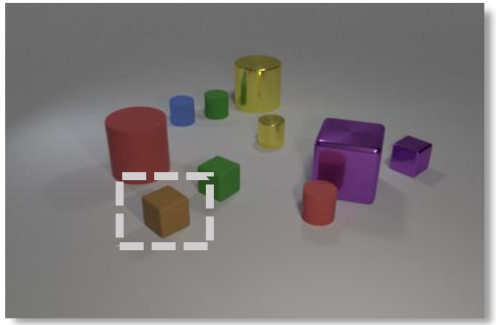
Image Difference Captioning

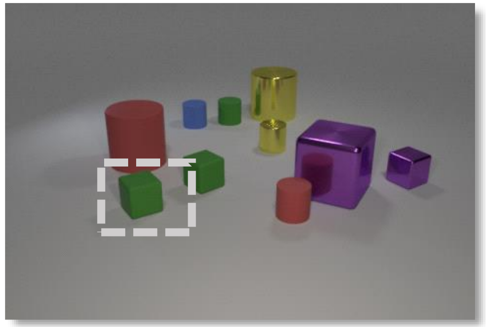
Perception





Comparison





Description

*" The **brown** matte cube changed to **green**. "*

Challenges

Challenge 1. Fine-grained Comprehension

e.g. differences lie in the tiny body parts of bird species (“feather” and “chest”)

Challenge 2. High-cost Annotation

data format is triplet *(img1, img2, description)*
existing manually annotated benchmark datasets are limited in data size

Our Motivation

We propose a **new pre-training and fine-tuning schema** for image difference captioning.

Challenge 1. Fine-grained Comprehension

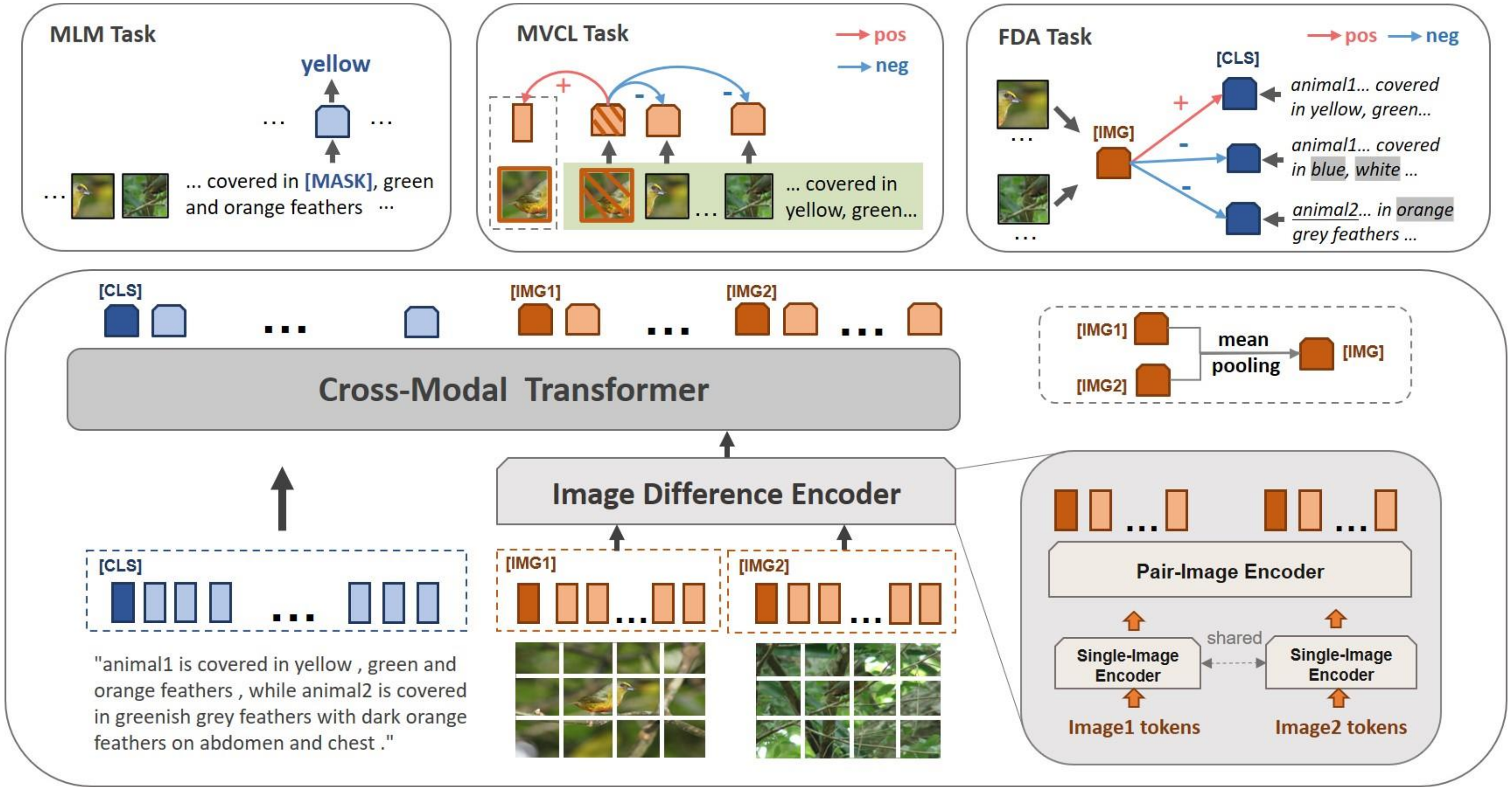
We design three self-supervised tasks to enhance the fine-grained cross-modal alignment by contrastive learning

Challenge 2. High-cost Annotation

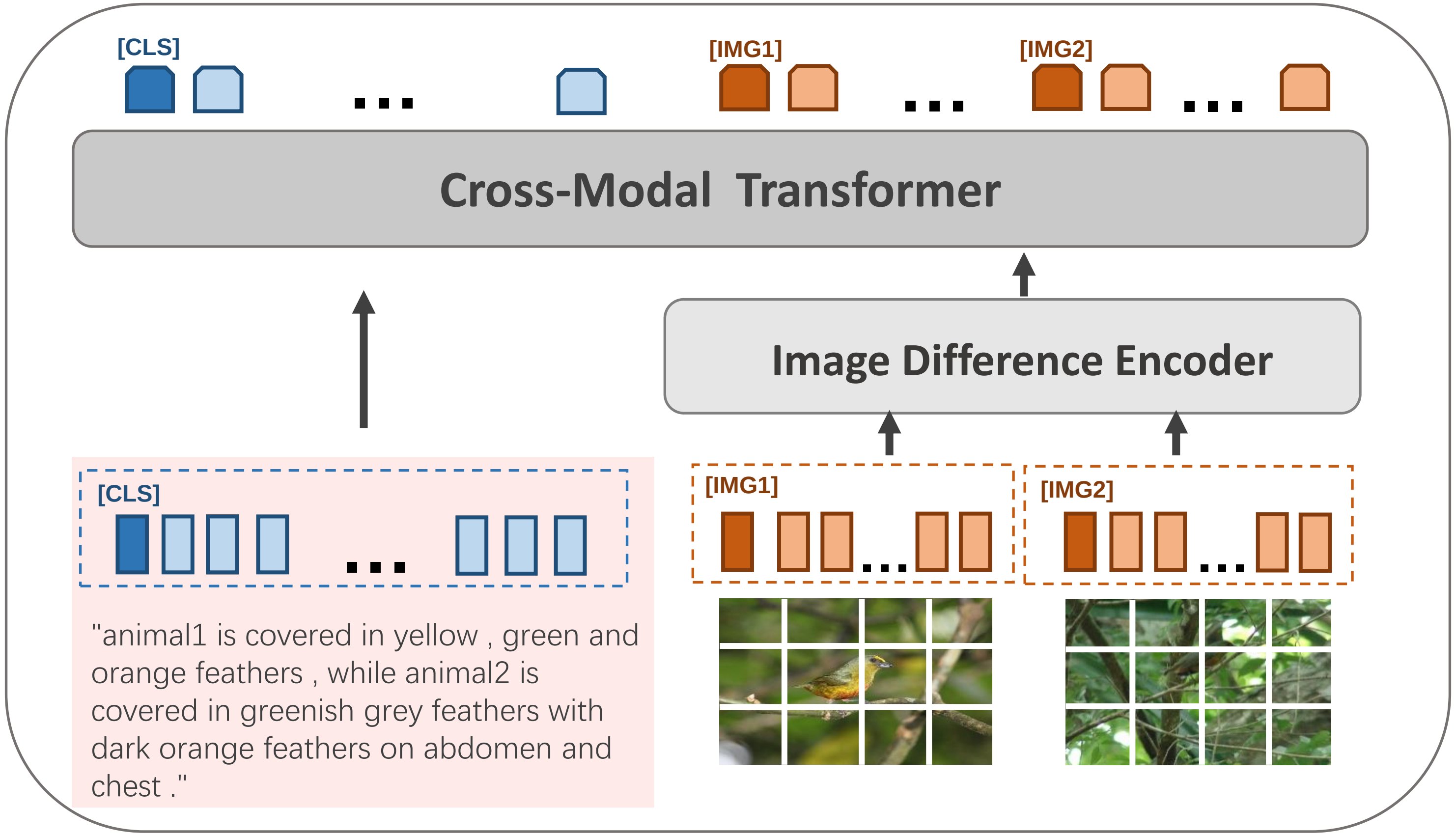
We use extra cross-task in-domain data in our framework to provide additional background knowledge

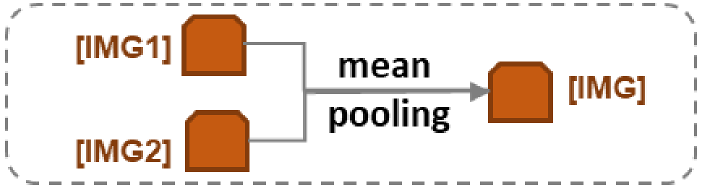
Method

We propose a new pre-training and fine-tuning paradigm for IDC with three pre-training tasks: MLM, MVCL and FDA.



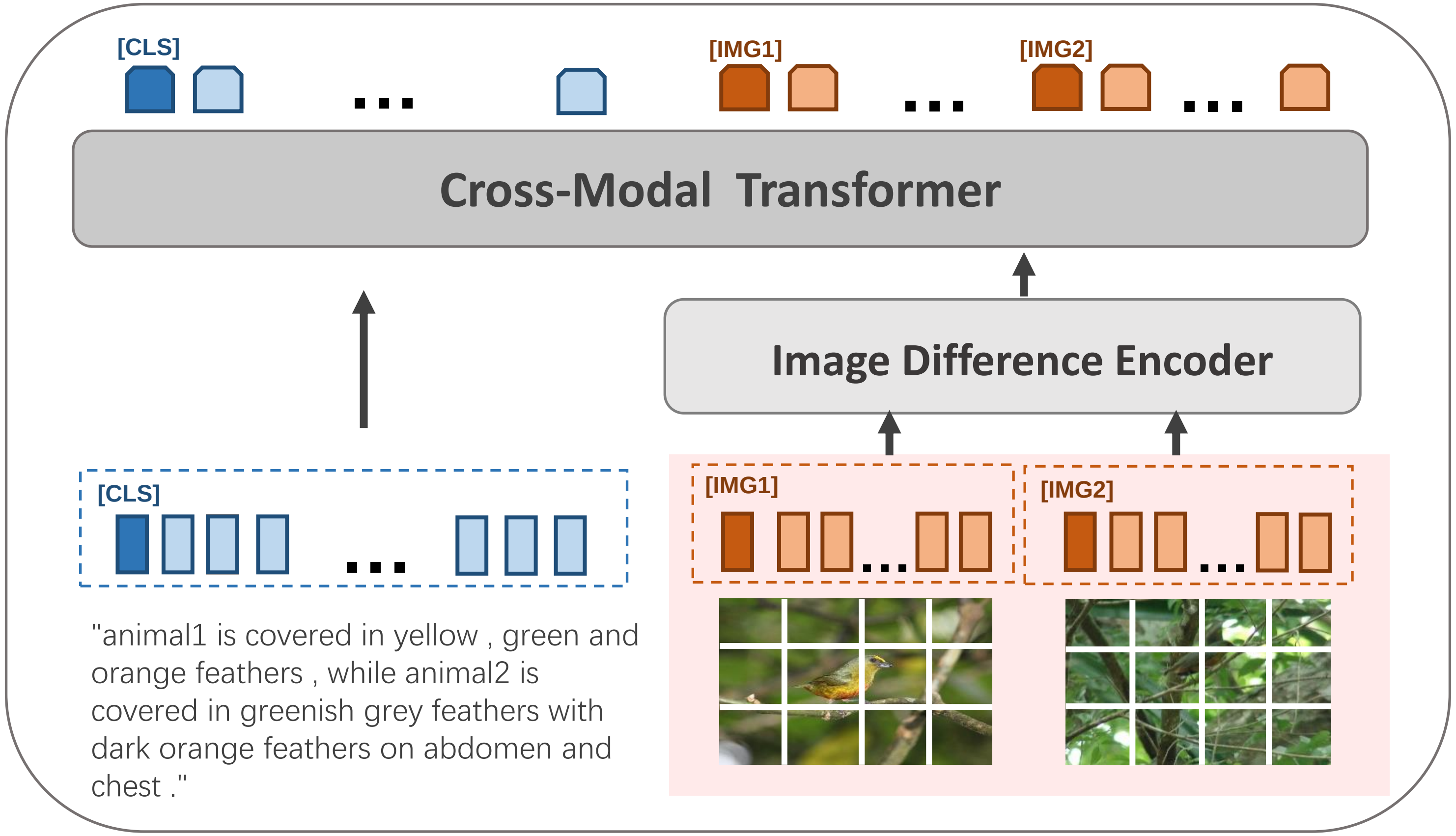
Input Representation

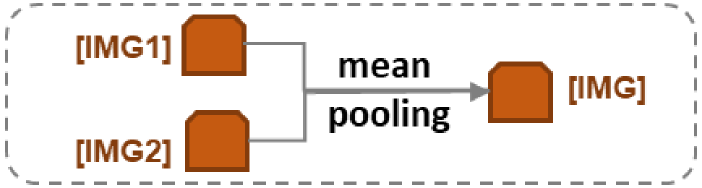




$$T = \{[\mathrm{CLS}], [\mathrm{BOS}], w_0, \ldots, w_M, [\mathrm{EOS}]\}$$

Input Representation

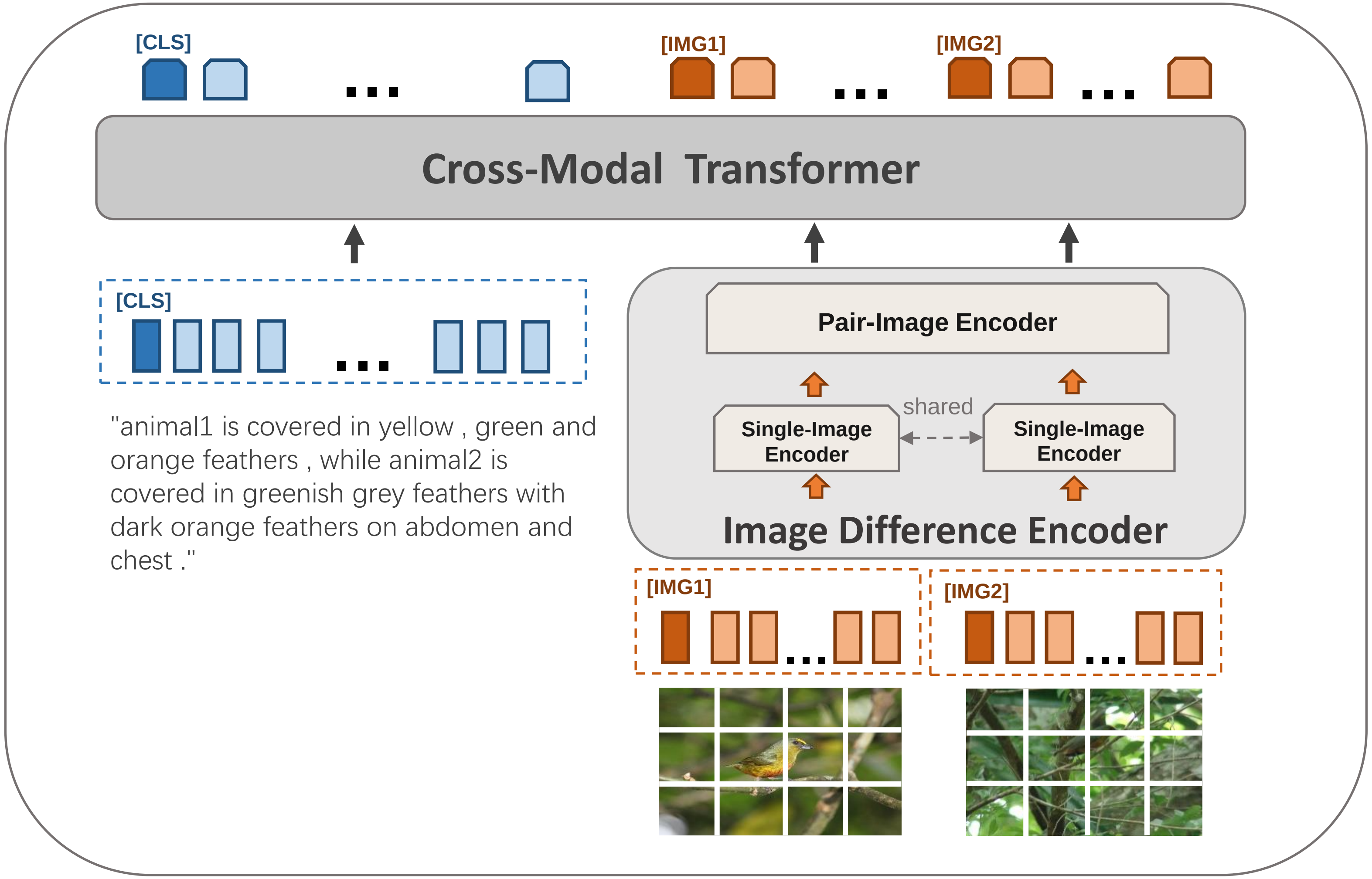




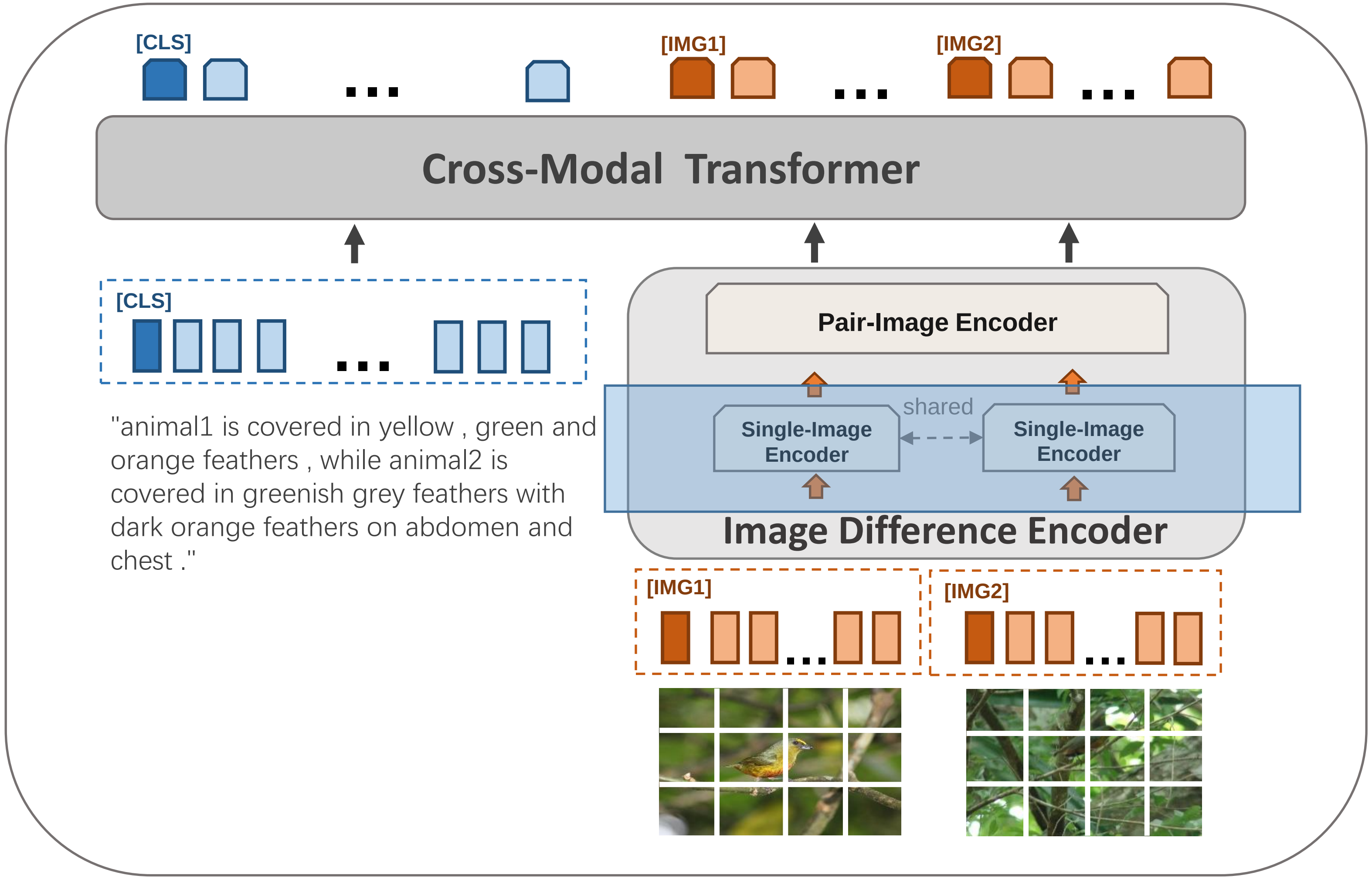
Input Embeddings
+ Positional Embeddings
+ Type Embeddings

$$V^{(1)} = \{[\text{IMG1}], v_0^{(1)}, \ldots, v_i^{(1)}, \ldots, v_N^{(1)}\}$$
$$V^{(2)} = \{[\text{IMG2}], v_0^{(2)}, \ldots, v_i^{(2)}, \ldots, v_N^{(2)}\}$$

Model Architecture

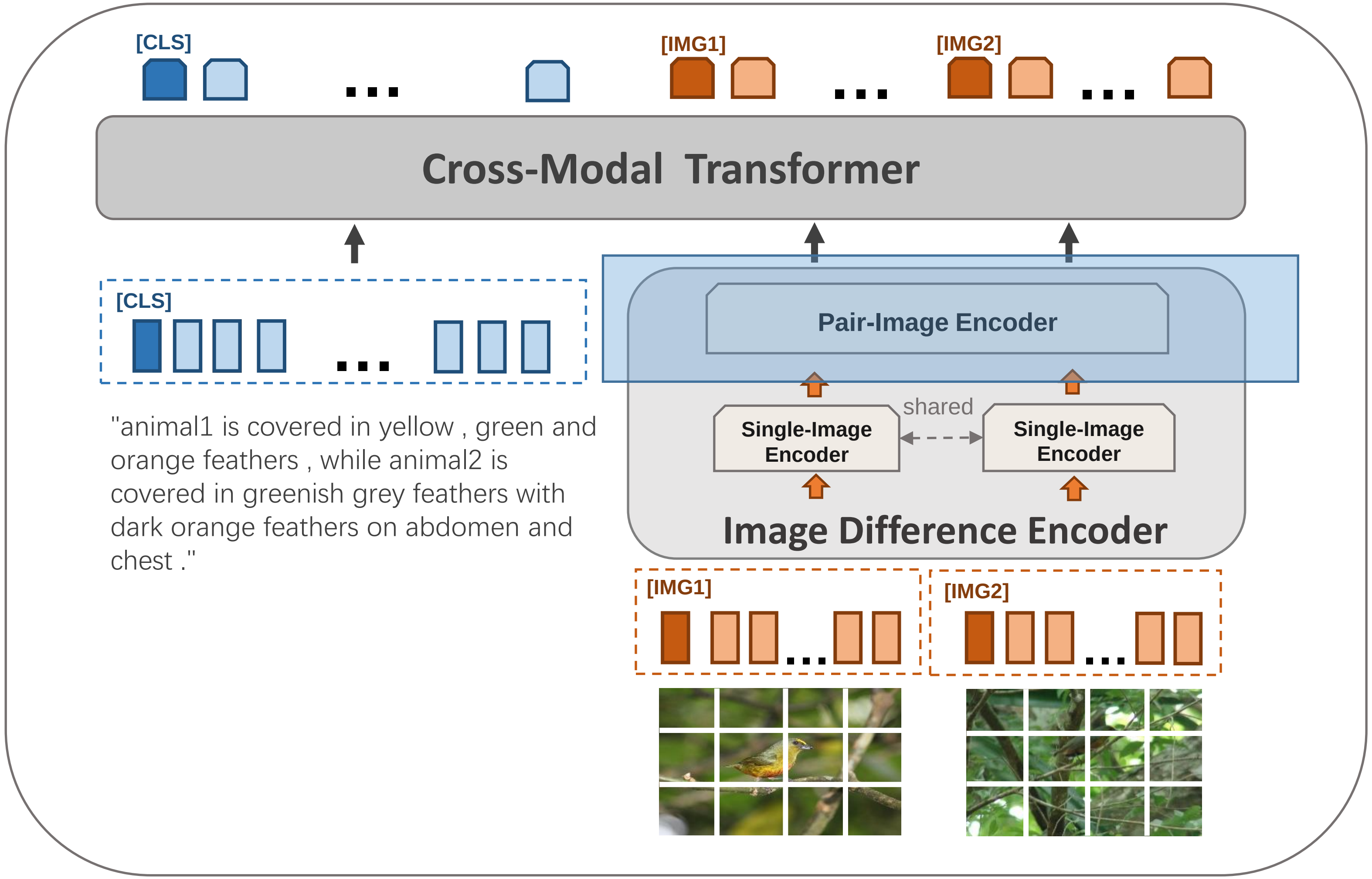


Model Architecture



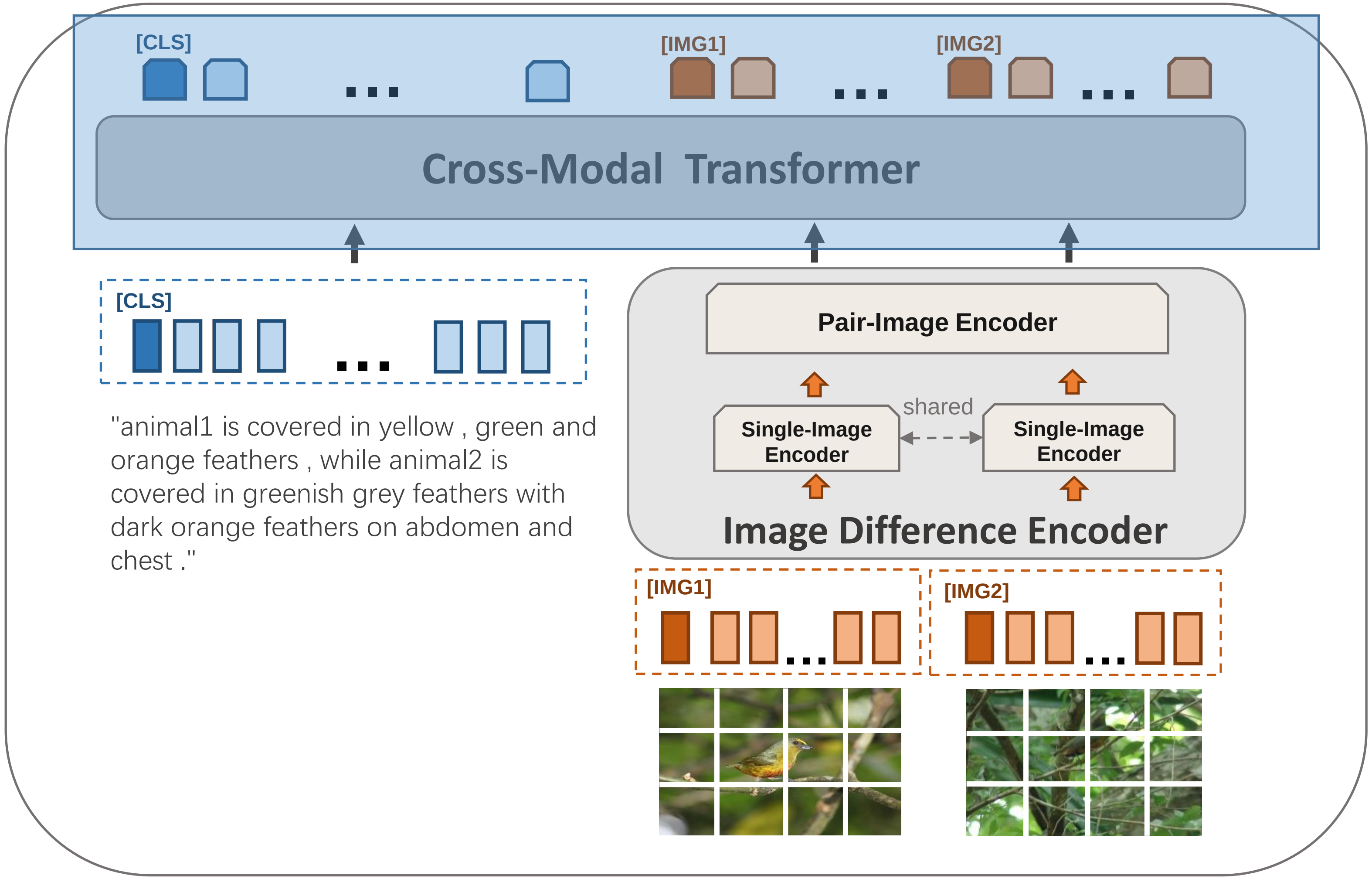
Perception

Model Architecture



Comparison

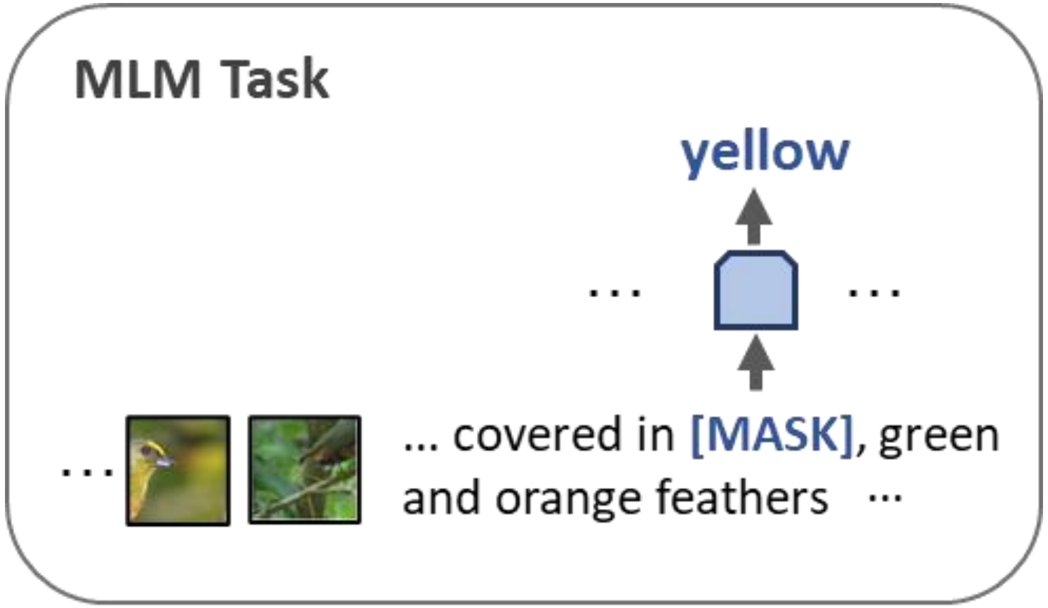
Model Architecture



Description

Pre-training Tasks

① Masked Language Modeling (MLM)

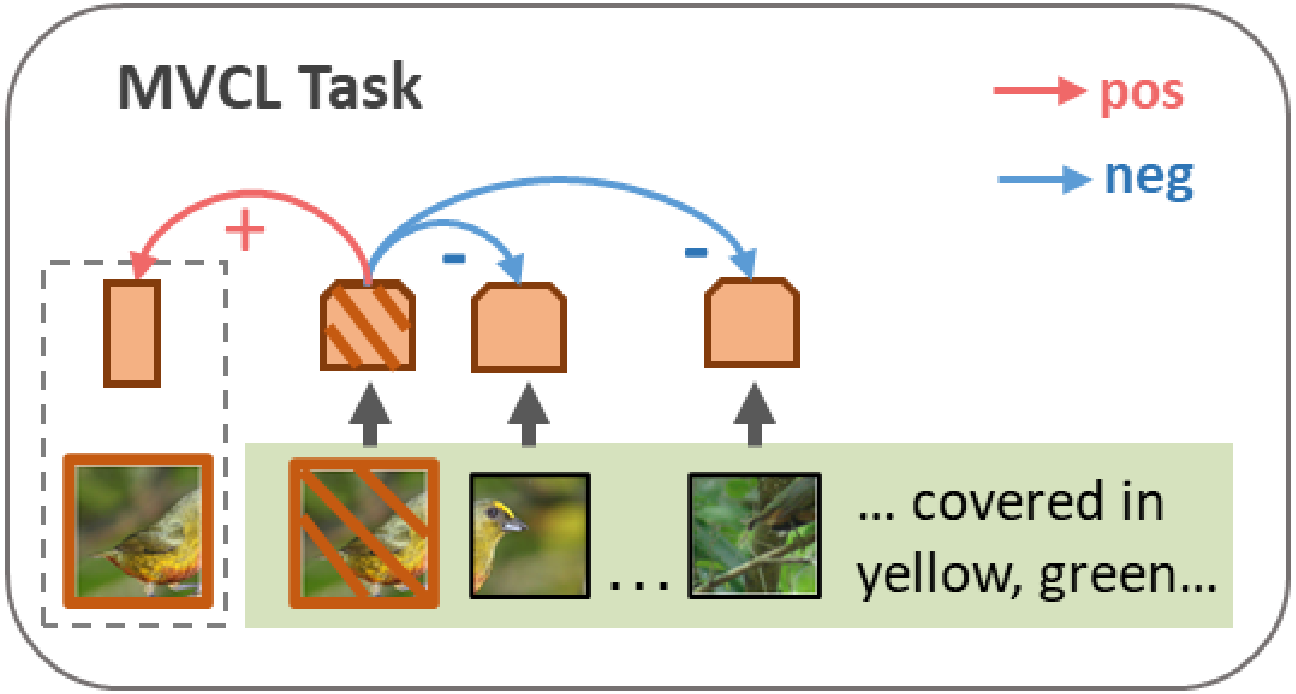


$$\mathcal{L}_{\mathrm{MLM}} = \mathbb{E}_{V,T\in D}\left[-\log P_\theta\left(w_m \mid w_{\setminus m}, \widetilde{V}^{(1)}, \widetilde{V}^{(2)}\right)\right]$$

Pre-training Tasks

② Masked Visual Contrastive Learning (MVCL)

Positive examples: the original feature before masking



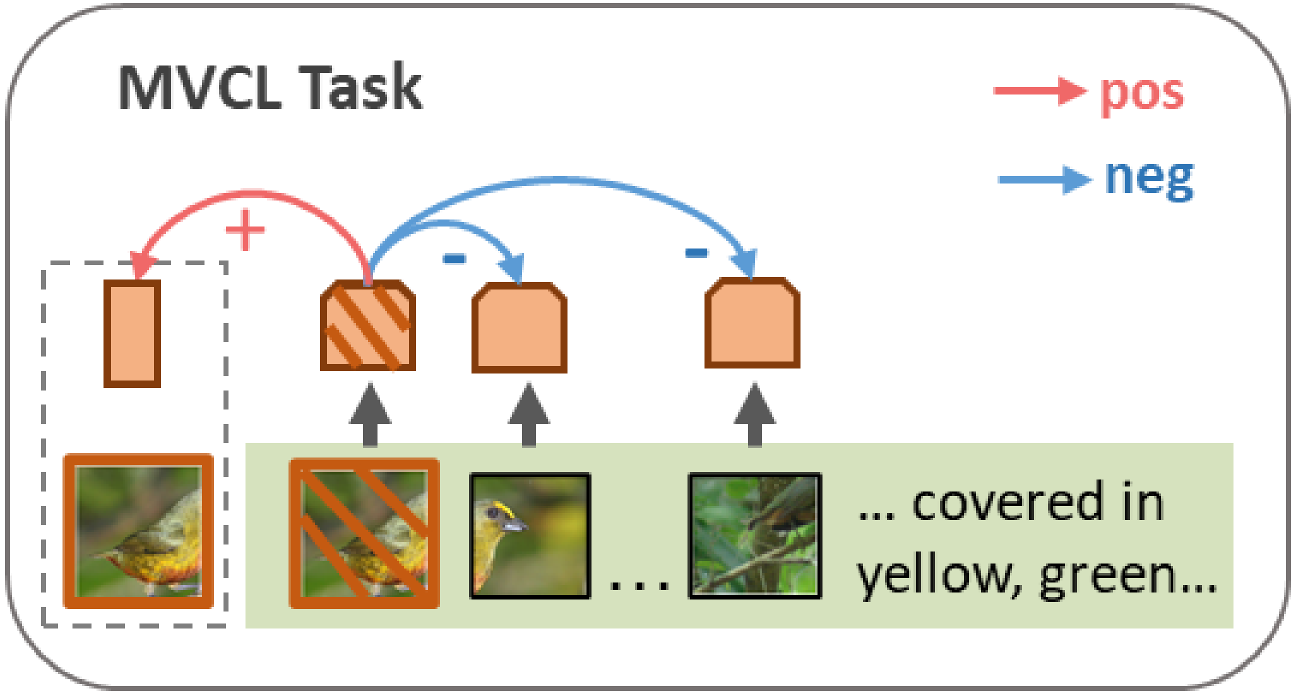
Negative examples: unmasked image features in the batch

$$\mathcal{L}_{\text{MVCL}} = \mathbb{E}_{V,T\in D}\left[-\log \frac{\exp\left(d(v_m, v_m^+)/\tau_1\right)}{\exp\left(d(v_m, v_m^+)/\tau_1\right) + \sum_{v'\in\mathcal{N}(v_m)} \exp\left(d(v_m, v')/\tau_1\right)}\right]$$

Pre-training Tasks

② Masked Visual Contrastive Learning (MVCL)

Positive examples:
the original feature before masking

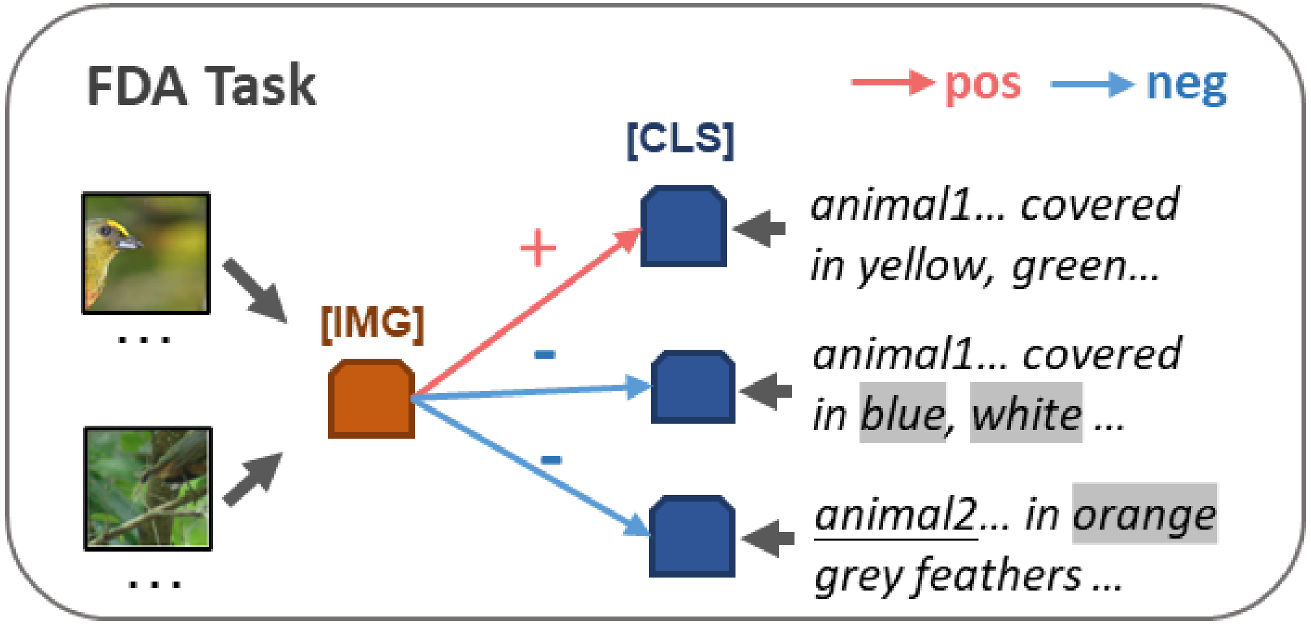


Negative examples:
unmasked image features in the batch

(, img2, description) or *(img1, , description)*

Pre-training Tasks

③ Fine-grained Difference Aligning (FDA)



$$\mathcal{L}_{\text{FDA}} = \mathbb{E}_{V,T\in D}\left[-\log \frac{\exp\left(d\left(V,T^{+}\right)/\tau_2\right)}{\exp\left(d\left(V,T^{+}\right)/\tau_2\right)+\sum_{T^{-}\in\mathcal{N}_T}\exp\left(d\left(V,T^{-}\right)/\tau_2\right)}\right]$$

Pre-training Tasks

- Construct hard negative samples by rewriting the original difference caption in three ways: Retrieve, Replace, Confuse

Original animal1 is brown with white tuft while animal2 is orange

Retrieve animal1 is brown with white tuft while animal2 is dark brown with grey tuft

Replace selected words [tuft, orange, brown]
animal1 is stocky with white spotting while animal2 is greenish

Confuse animal2 is brown with white tuft while animal1 is orange

Finetuning and Inference

Finetuning

MLM + uni-directional attention mask

Inference

generates the difference caption word by word
based on visual difference semantics.

Data expansion strategy

Utilize extra **cross-task in-domain** data to provide additional background knowledge.

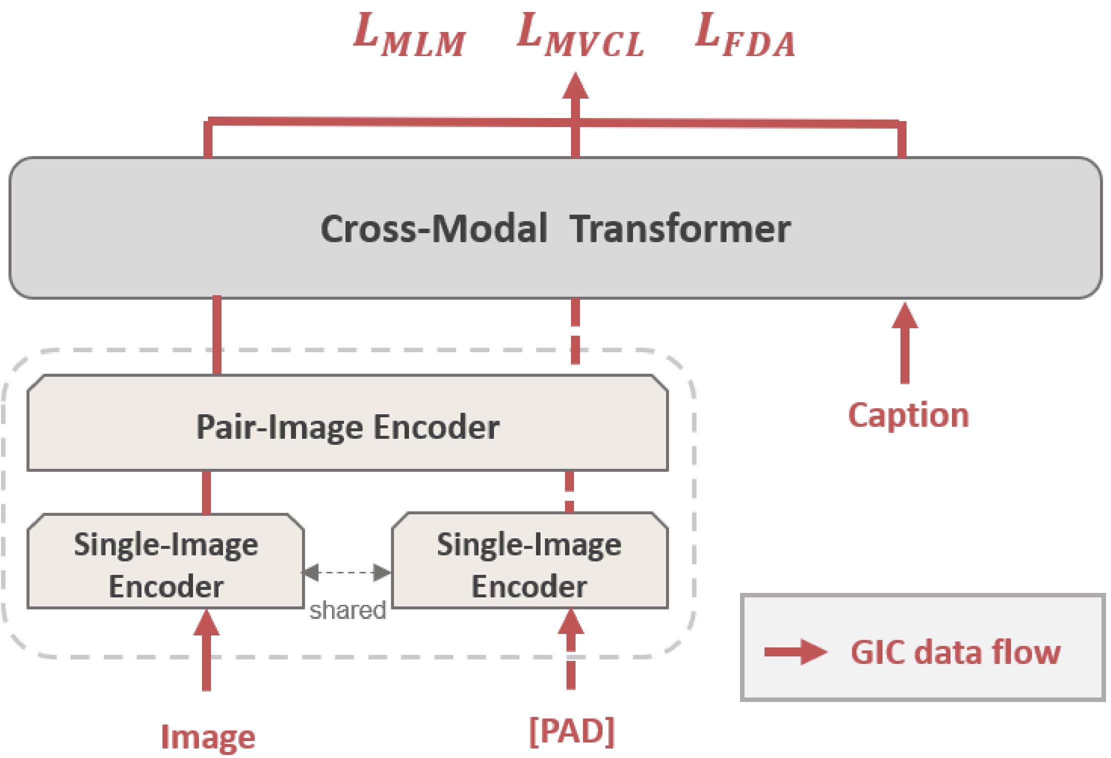
- **General image captioning(GIC)** data

- **Fine-grained visual classification(FGVC)** data

Data expansion strategy

- The GIC data is in the (image, text) format, which can facilitate the model to learn preliminary cross-modal alignment.

"this bird has gray feathers with a white throat, breast, and abdomen."

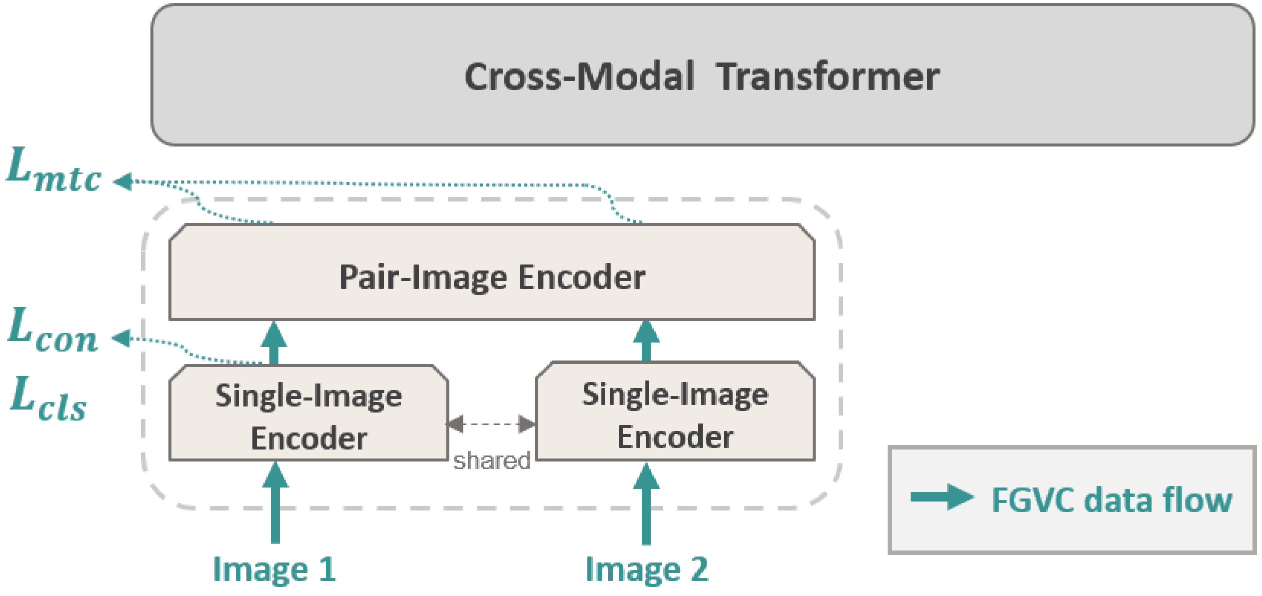


Data expansion strategy

- The FGVC data is in the (image, class label) format. It can enhance image difference encoder to learn more discriminative visual representations.

Ring-necked Duck

Greater Scaup

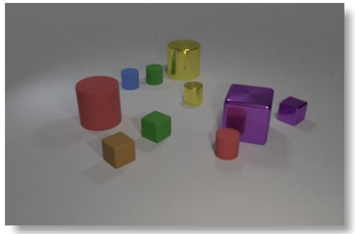


L_{mtc}: predict whether two images are from one class (0/1)

L_{cls} / L_{con}: predict the class label of an image

Experiments

Benchmark Datasets



- CLEVR-Change dataset
  It has 67,660, 3,976 and 7,970 image pairs for training, validation and test split respectively. Each image pair is annotated with 6.2 captions on average.

- Birds-to-Words dataset
  It has 4,860 image pairs and each pair corresponds to 3.31 annotated captions on average.

Metrics

standard image captioning metrics including BLEU, METEOR, ROUGE-L and CIDEr(CIDEr-D)

Comparison with SOTA

Results on Birds-to-Words

- B4, M, R, and C(D) are short for BLEU-4, METEOR, ROUGE-L and CIDEr(D).
- The main metric **ROUGE-L** on this dataset is highlighted.

Model	B4	M	C(D)	**R**
Neural Naturalist (2019)	22.0	-	25.0	43.0
Relational Speaker (2019)	21.5	22.4	5.8	43.4
DUDA (2019)	23.9	21.9	4.6	44.3
L2C (2021)	31.3	-	15.1	45.3
L2C(+CUB) (2021)	**31.8**	-	16.3	45.6
Ours	28.0	23.1	18.6	48.4
Ours(+Extra Data)	31.0	**23.4**	**25.3**	**49.1**

Comparison with SOTA

Results on CLEVR-Change

- B4, M, R, and C are short for BLEU-4, METEOR, ROUGE-L and CIDEr.
- The main metric **CIDEr** on this dataset is highlighted.

Model	B4	M	R	C
Capt-Dual-Att (2019)	43.5	32.7	-	108.5
DUDA (2019)	47.3	33.9	-	112.0
VAM (2020)	50.3	37.0	69.7	114.9
VAM+ (2020)	**51.3**	**37.8**	70.4	115.8
IFDC (2021a)	49.2	32.5	69.1	118.7
DUDA+Aux (2021)	51.2	37.7	70.5	115.4
Ours	51.2	36.2	**71.7**	**128.9**

Ablation Study

- **DE** is short for image **D**ifference **E**ncoder
- B4, M, R, and C are short for BLEU-4, METEOR, ROUGE-L and CIDEr.

Pre-training Tasks	DE	B4	M	R	C
1 None	✓	32.7	27.7	57.2	89.8
2 MLM	✓	36.7	28.2	60.9	94.9
3 MLM + MVCL	✓	50.3	37.6	70.6	119.7
4 MLM + MVCL + FDA	✓	51.2	36.2	71.7	128.9
5 MLM + MVCL + FDA	✗	49.2	35.8	68.8	107.9
6 w/o Distractor Judging	✓	49.8	36.9	69.2	123.5

Cross-task Data Usage

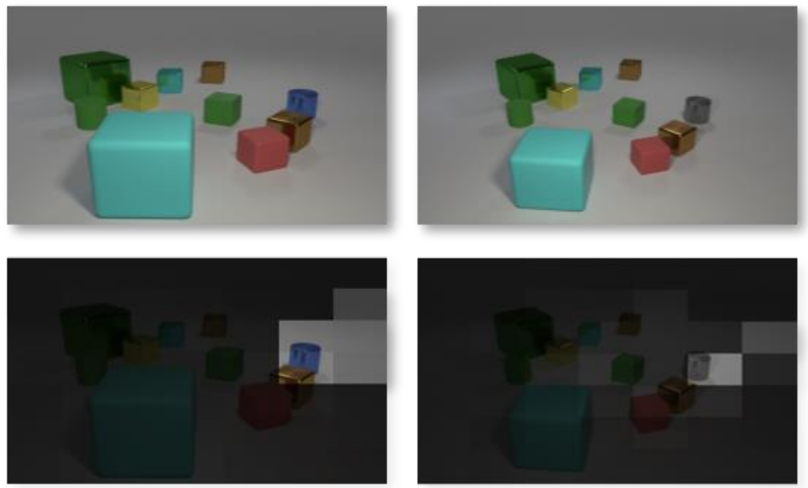
- Birds-to-Words(B2W): an image difference captioning dataset
- CUB: a general image captioning dataset
- NABirds(NAB): a fine-grained visual classification dataset

Model	B2W	CUB	NAB	B4	M	C(D)	R
L2C	✓			31.3	-	15.1	45.3
	✓	✓		31.8	-	16.3	45.6
Ours	✓			28.0	23.1	18.6	48.4
	✓	✓		29.3	23.1	23.8	48.5
	✓		✓	27.5	23.3	21.9	48.5
	✓	✓	✓	31.0	23.4	25.3	49.1

Case Visualization

Semantic Change

(a)



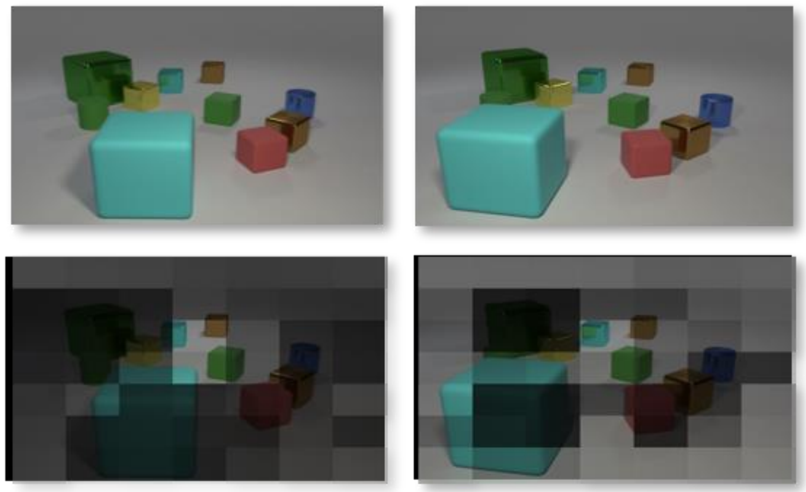
Ours: *the **small blue metal cylinder** that is to the right of the small yellow thing **became gray***

DUDA: *the small green metal cylinder that is behind the small brown matte cylinder is missing*

GT: *the blue metallic thing became gray*

Distractors

(b)



Ours: *the scene is the same as before*

DUDA: *the scene is the same as before*

GT: *the two scenes seem identical*

Distractors: only non-semantic differences between the images (e.g. angle, zoom, or illumination changes)

Case Visualization

(c)

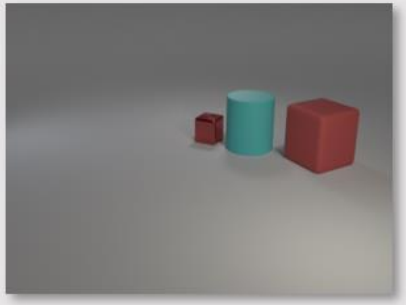
Ours: *animal1 has* ***red feathers on its head*** *, and* ***wings*** *and tail . animal2 has* ***a brown head*** *. animal2 has* ***a brown*** *and white* ***breast*** *.*

Neural Naturalist: *animal1 has a red head . animal2 has a brown head .*

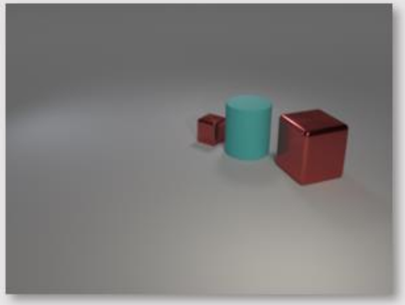
GT: *animal1 has a red beak , while animal2 has a pale grey beak . animal1 ' s vivid coloring includes red , violet , tan , rust , blue , and brown . in contrast , animal2 ' s coloring is mostly yellow and dark brown . animal1 has black legs , while animal2 has red legs .*

Visualization of Cross-modal Alignment

An unseen triplet sample from the test set



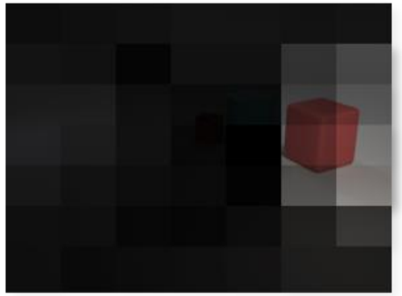
[img1]



[img2]

the ***large red matte cube*** that is on the ***right*** side of the ***tiny red metallic block*** turned ***metallic***

large red matte cube

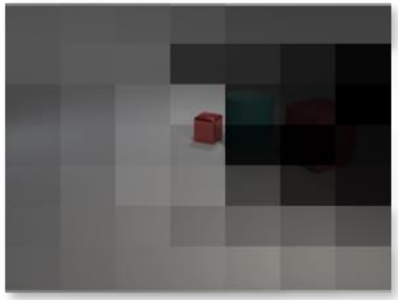


[img1]

right

[img1]

tiny red metallic block



[img1]

metallic

[img2]

Conclusions

- **New schema**

a pre-training and finetuning paradigm for IDC task

- **New pre-training tasks**

propose MLM. MVCL, FDA tasks with contrastive learning to enhance fine-grained cross-modal alignment

- **Cross-task data expansion**

utilize GIC and FGVC datasets to provide additional in-domain knowledge

Thank You!

If any questions, feel free to contact
{linliyao, wy.wang, qjin}@ruc.edu.cn