# Ranking-Aware Adapter for Text-Driven Image Ordering with CLIP

Wei-Hsiang Yu[1] Yen-Yu Lin[1] Ming-Hsuan Yang[2] Yi-Hsuan Tsai[3]

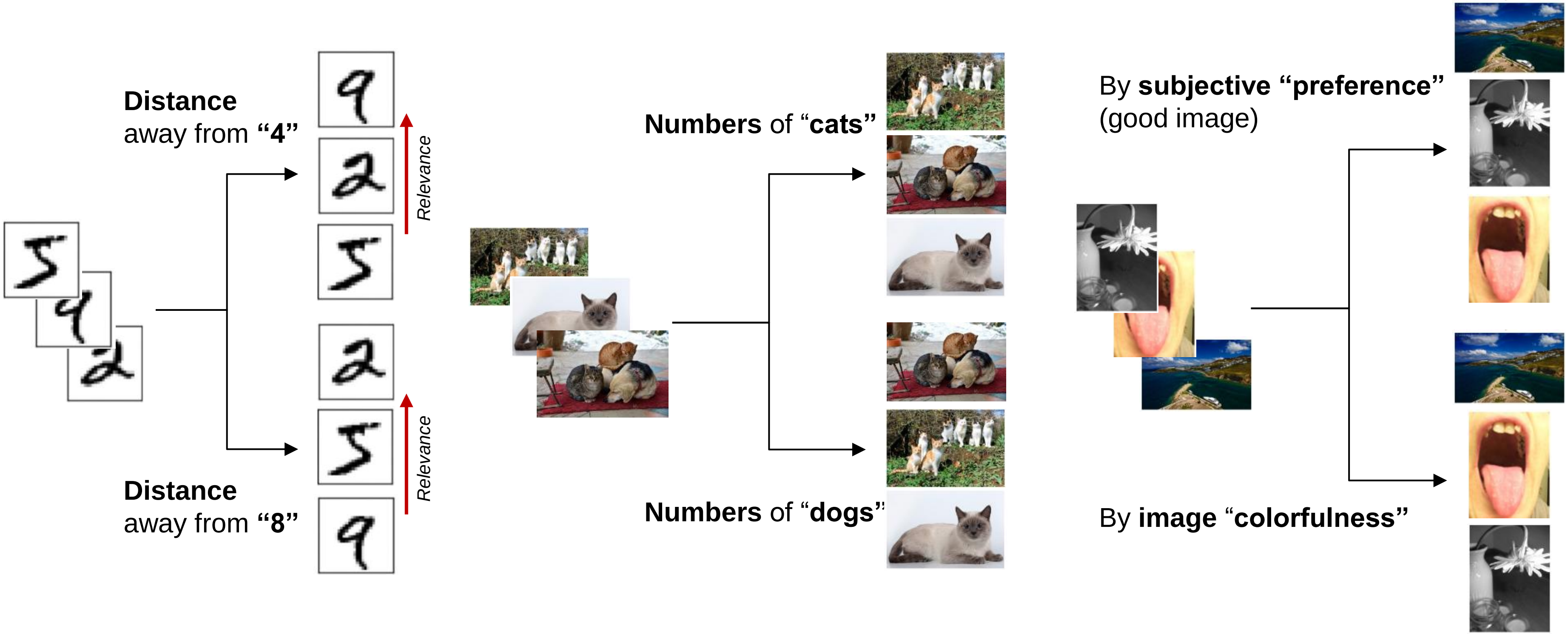
[1]National Yang Ming Chiao Tung University [2]UC Merced [3]Atmanity Inc.

Website

# Ranking-Aware Adapter for Text-Driven Image Ordering with CLIP

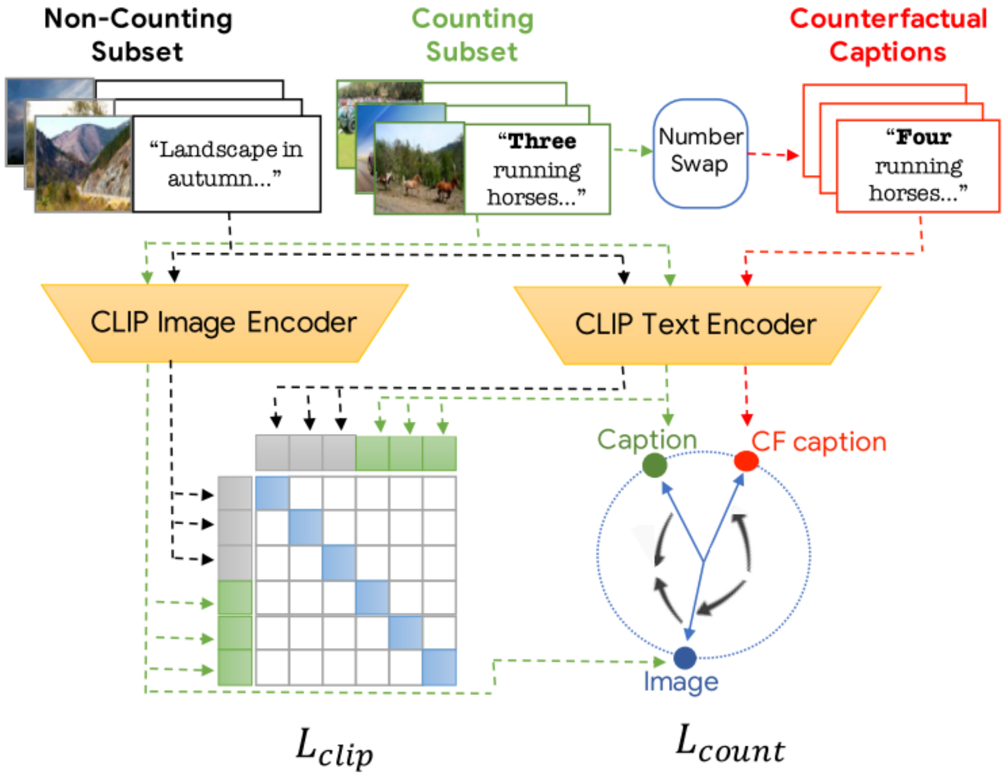
# Problem formulation

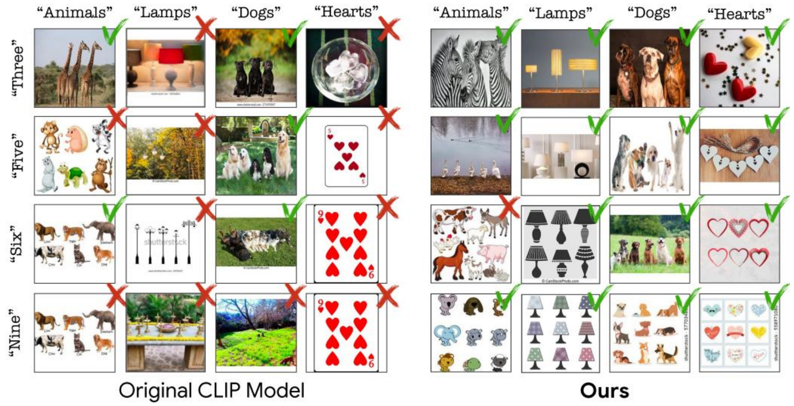
- Arrange a set of images according to user requirements.



# Previous efforts to text-driven image ordering

- Re-align “Number” to image using contrastive learning

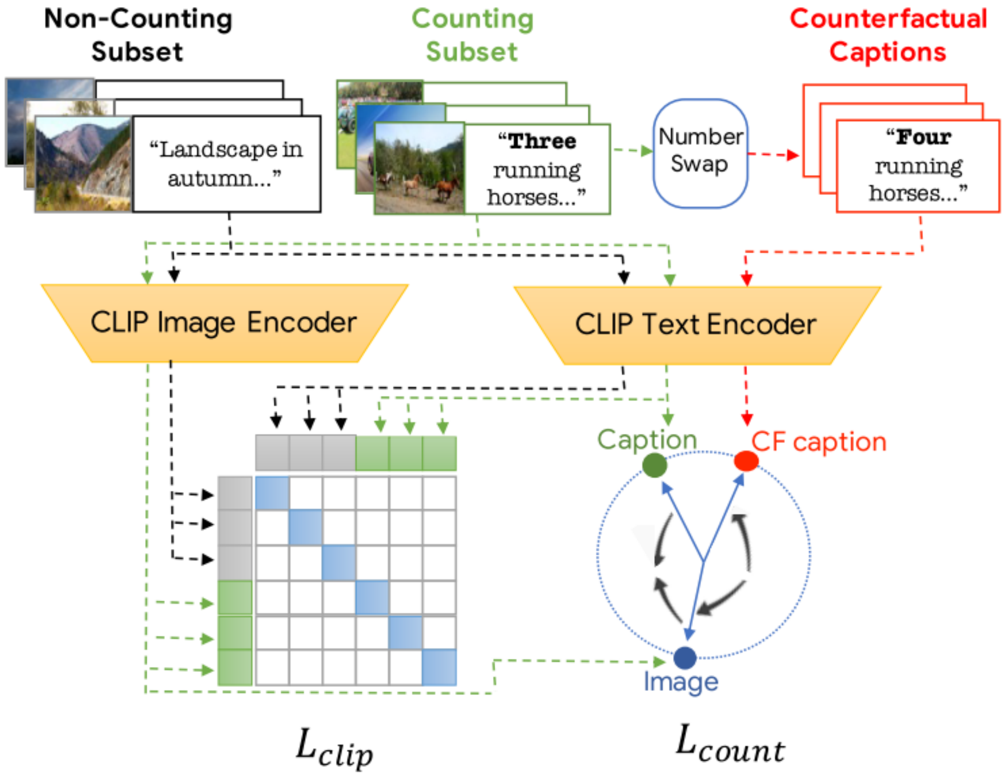



***Teach CLIP to Count to Ten, ICCV’23***

# Previous efforts to text-driven image ordering

## - Re-align “Number” to image using contrastive learning



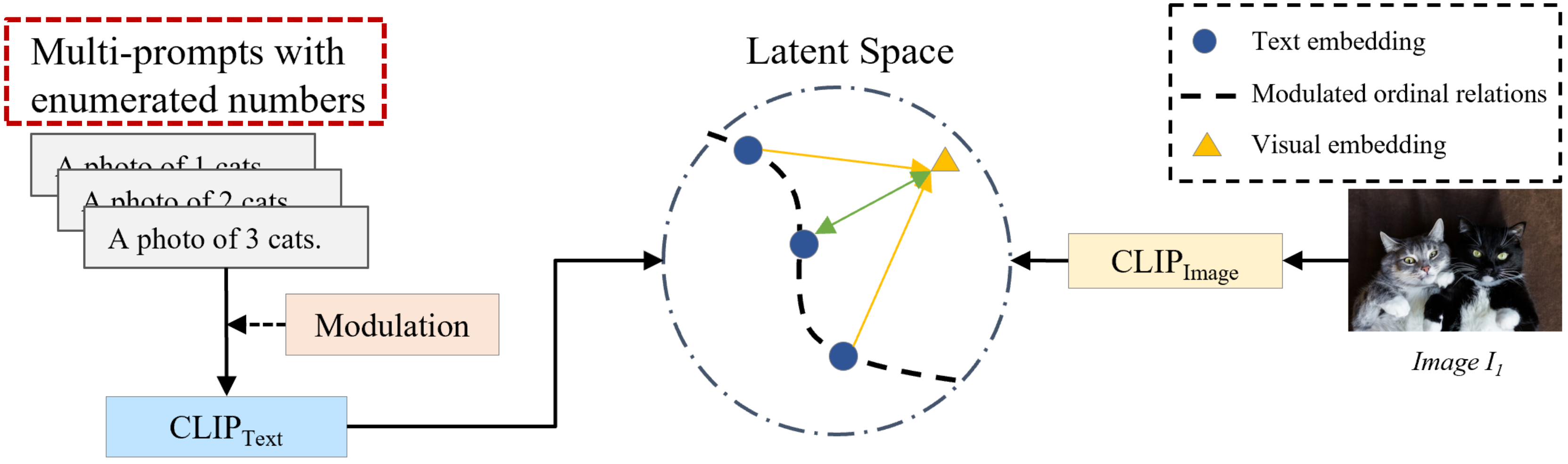

**Values span a wide range? (e.g., facial age)**

**Continuous and subjective values?**
**(e.g., image contrast, image aesthetics, etc.)**

***Teach CLIP to Count to Ten, ICCV’23***

# Motivation 1 – Contrastive learning with enumerated numbers

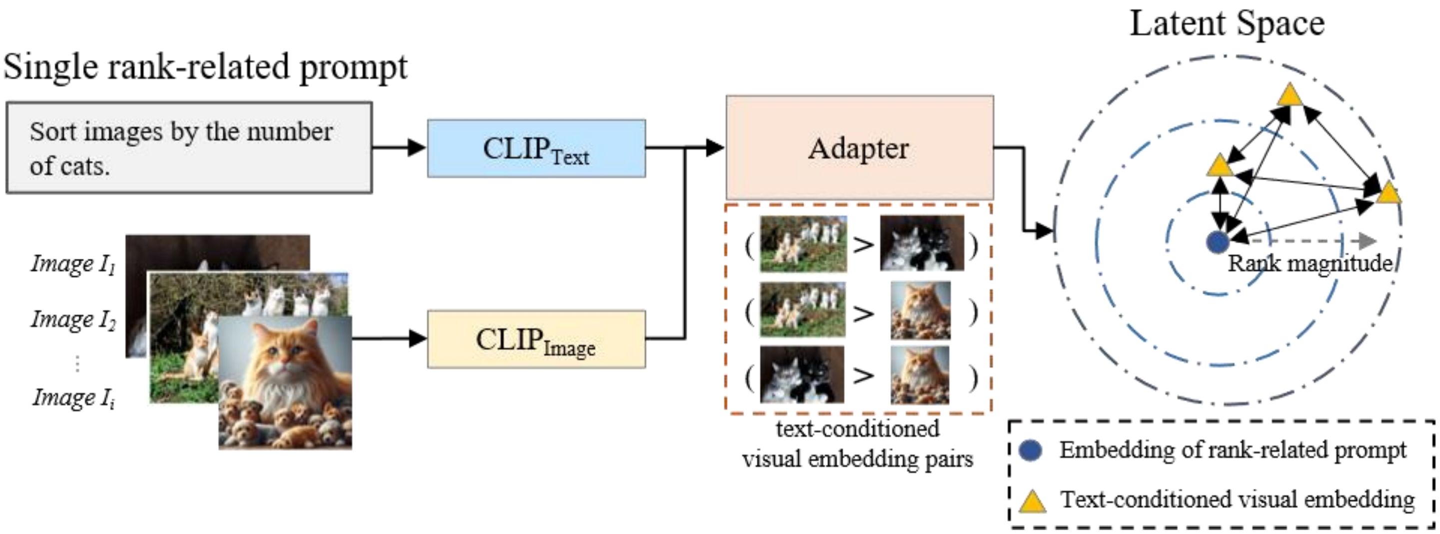
- Existing methods focus on aligning numerical text prompts to the image
  - Difficult to handle continuous values spanning a wide range - binning
  - Requires to enumerate all attribute-value combinations – inefficient




***Concept of prior works***

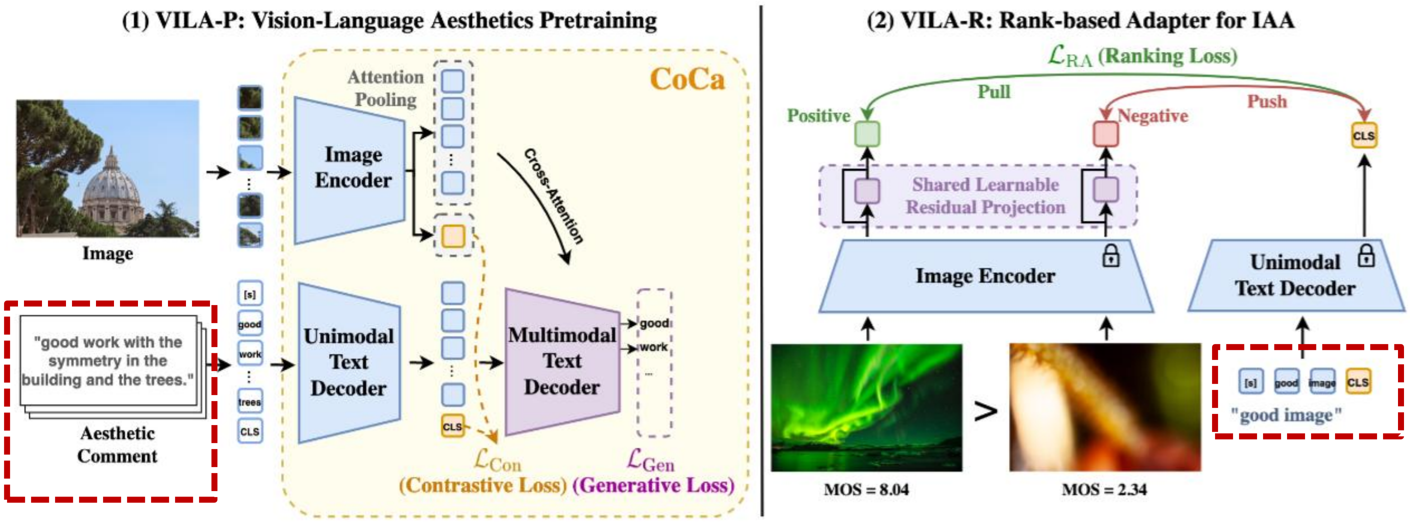
# Idea – Single rank-related prompt with image pairs

- Introduce a "single rank-related prompt" – The attribute used to sort images.
- Generate text-conditioned image embedding pairs and supervise using their numerical order with a learning-to-rank framework.



# Motivation 2 – Learning-to-Rank in Visual-Text Alignment

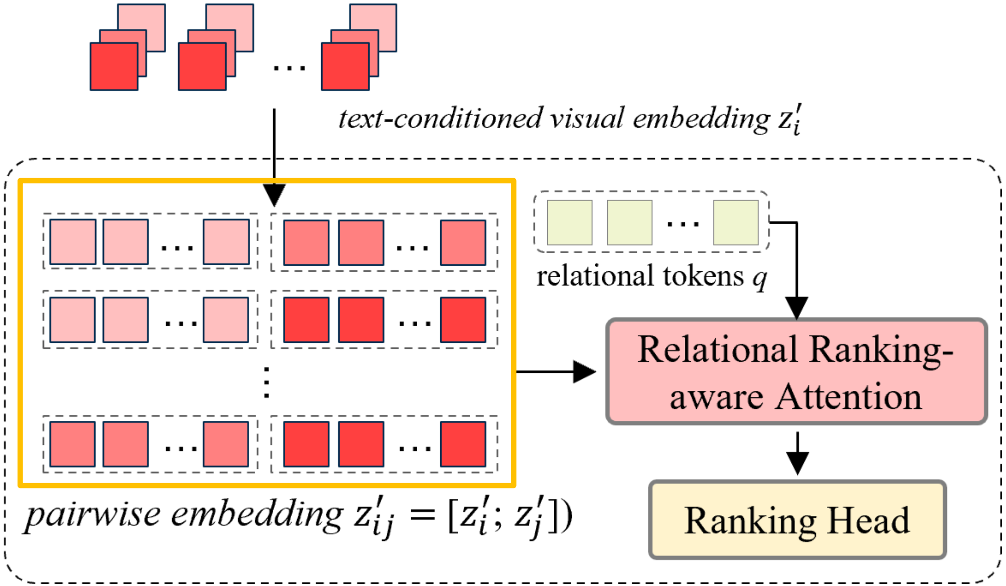
- Existing methods pretrain the text encoder on a task-specific dataset and then use a learning-to-rank framework to align images to specific attributes.
  - Requires fine-tuning the text encoder on a specific dataset (e.g., aesthetic comment).



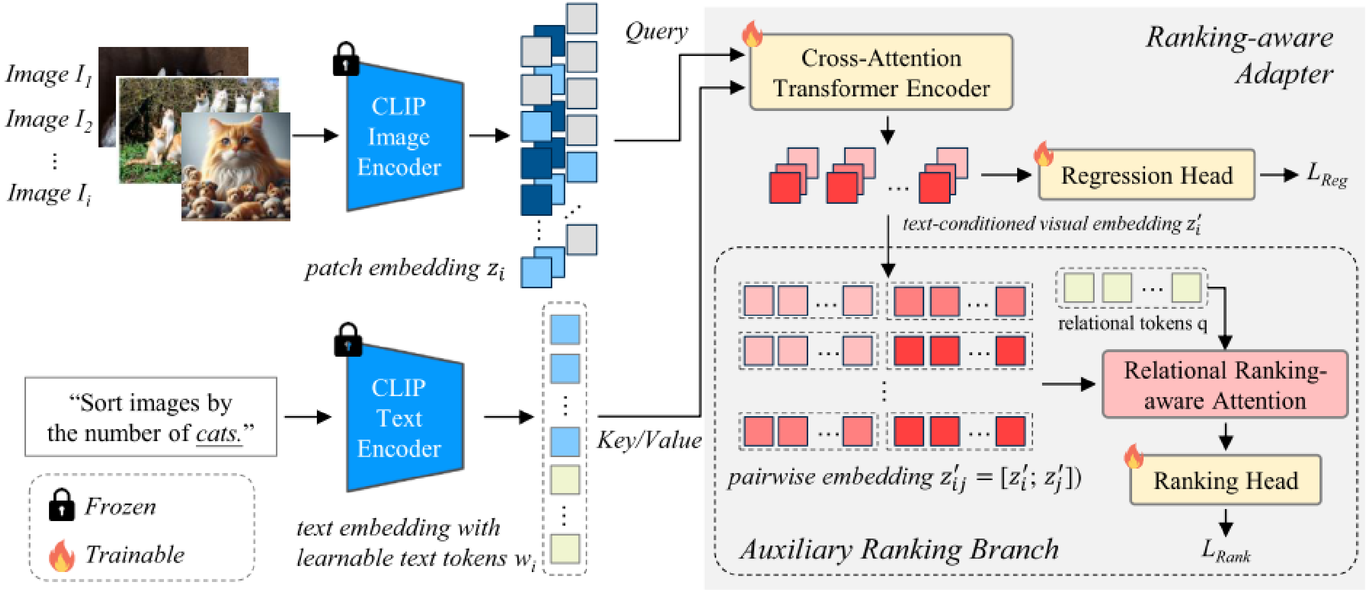

*"VILA", CVPR'23*

# Idea – Learning from text-conditioned visual distinctions

- Paired embedding difference ∝ queried value difference between the image pair
  - Design a relational ranking-aware attention module to extract the text-conditioned visual distinctions

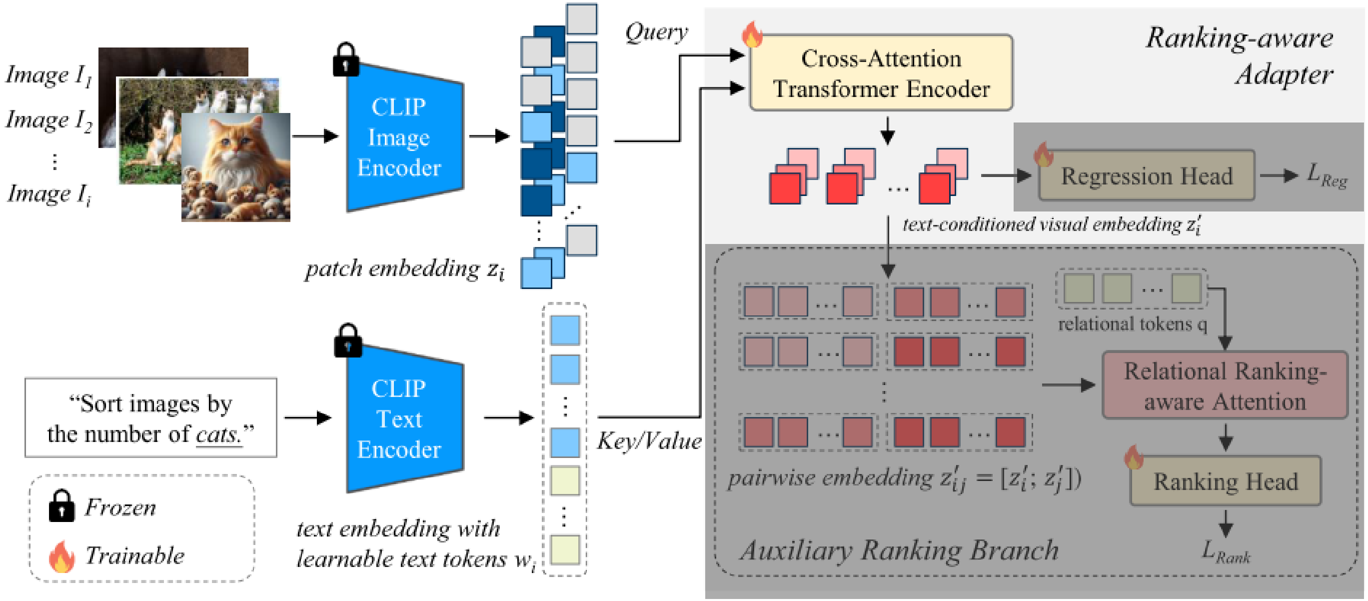


# Our Approach
- Ranking-Aware Adapter for Text-Driven Image Ordering with CLIP

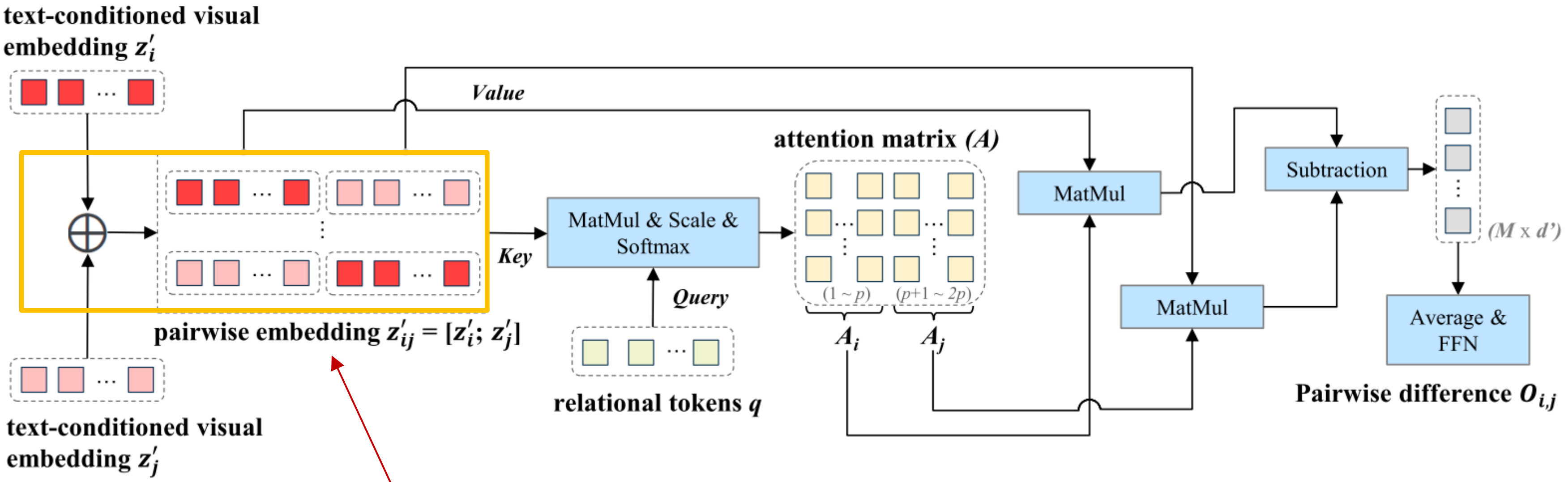


# Our Approach

- Ranking-Aware Adapter for Text-Driven Image Ordering with CLIP

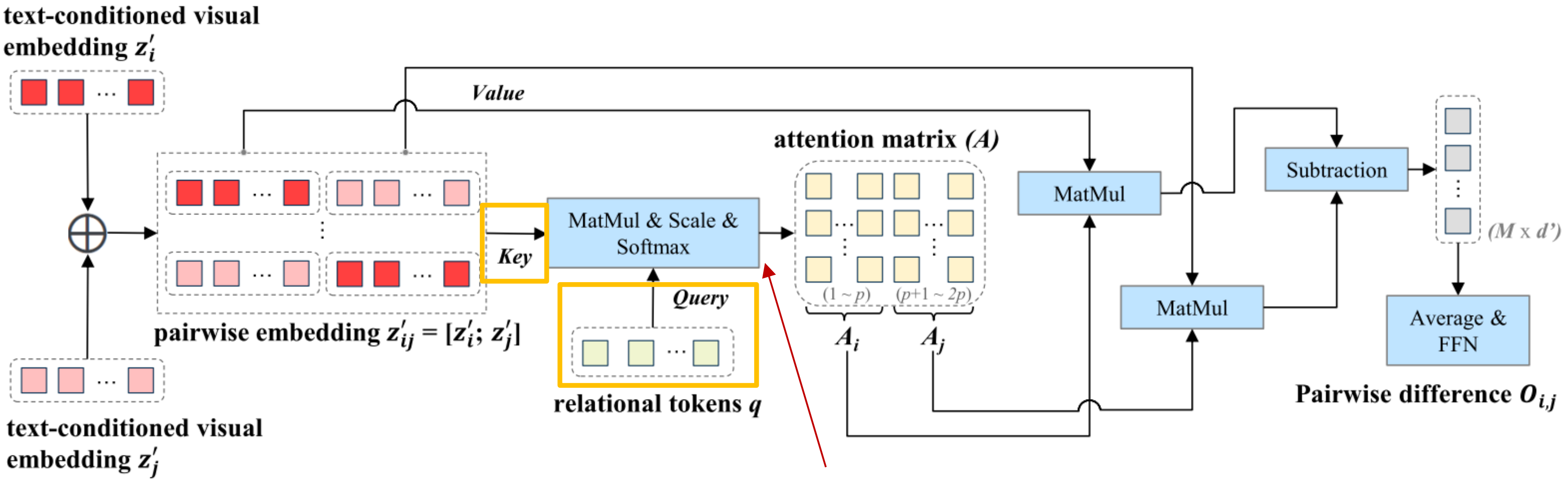


# Relational Ranking-aware Attention




Generate embedding pairs

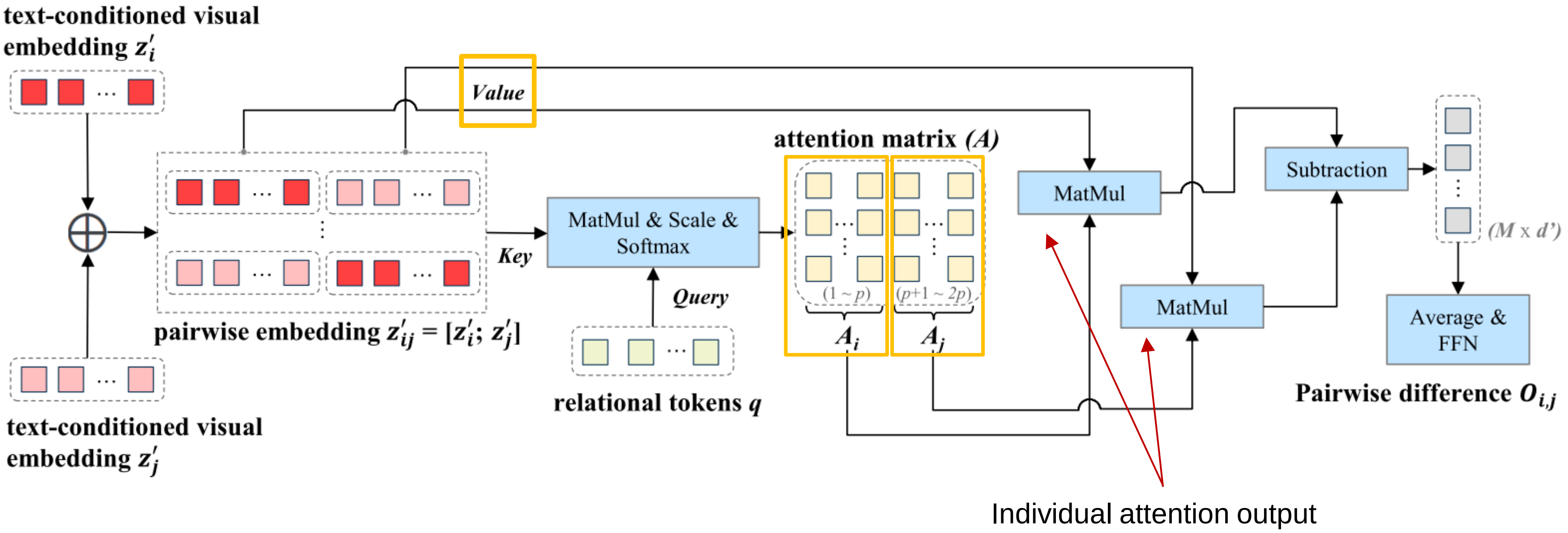
# Relational Ranking-aware Attention




Attention scores across a pair of embedding

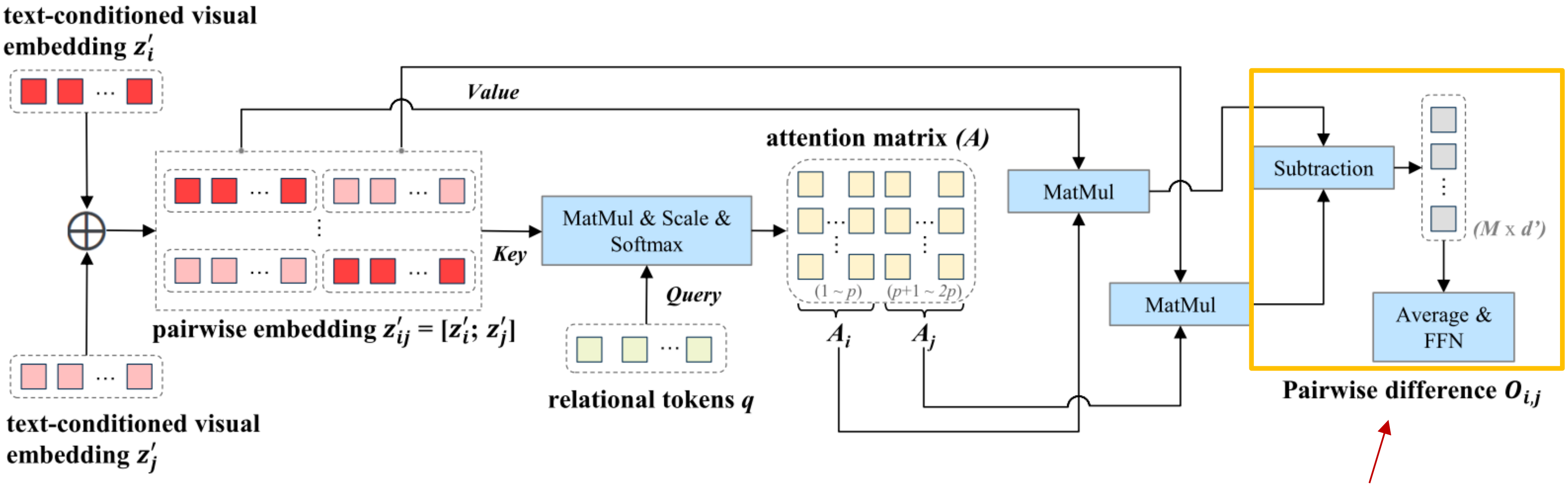
$$A = \mathtt{Softmax}\left(\frac{q \cdot (k_i \oplus k_j)^T}{\sqrt{d'}}\right) = \mathtt{Softmax}\left(\frac{q \cdot K^T}{\sqrt{d'}}\right)$$

# Relational Ranking-aware Attention




$$O_i = A_i \cdot V_i \quad \text{and} \quad O_j = A_j \cdot V_j.$$

# Relational Ranking-aware Attention

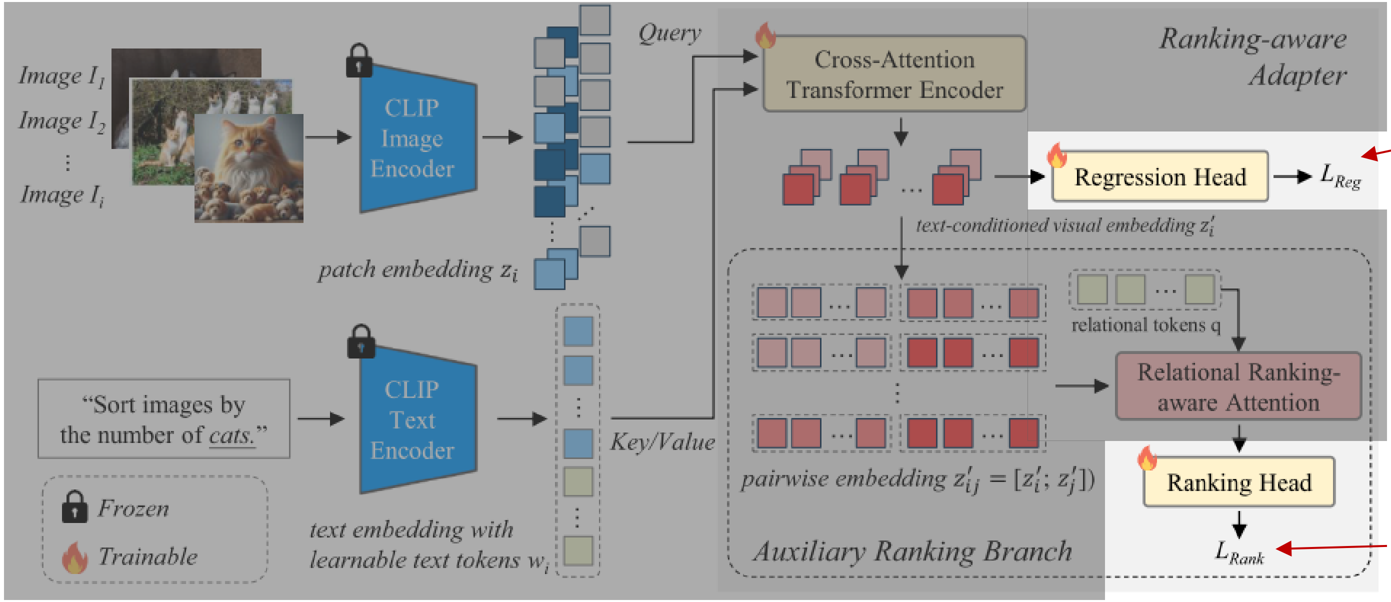



Enforce the relative difference

$$O_{i,j} = FFN\left(\frac{\sum_{m=1}^{M} \boxed{(O_{i,m} - O_{j,m})}}{M}\right)$$

# Our Approach

- Ranking-Aware Adapter for Text-Driven Image Ordering with CLIP




Predict image ranking score

$$L_{reg} = \begin{cases} \frac{1}{2}(y_i - s_i)^2, & \text{if } |y_i - s_i| < 1, \\ |y_i - s_i| - 0.5, & \text{otherwise.} \end{cases}$$

Encourage relative ordering

$$L_{rank}(y, O_{i,j}) = \sum_{\{(i,j)|y_i > y_j\}} \max(0, 1 - O_{i,j}).$$

# Experiment Settings

- Datasets
  - **Visual Quantity**: *COCO* (object count sorting)
  - **Visual Quality**: *KonIQA-10k* (image quality) and *AVA* (image aesthetics)
  - **Perceptual Concepts**: *Adience* (facial age) and *historical colored image* (photo taken decade)
- Evaluation Metrics
  - Object count sorting and IQA/IAA – **Pearson's correlation** and **Spearman's correlation**
  - Facial aging and historical colored image aging – **Mean absolute error** (MAE) and **Accuracy**

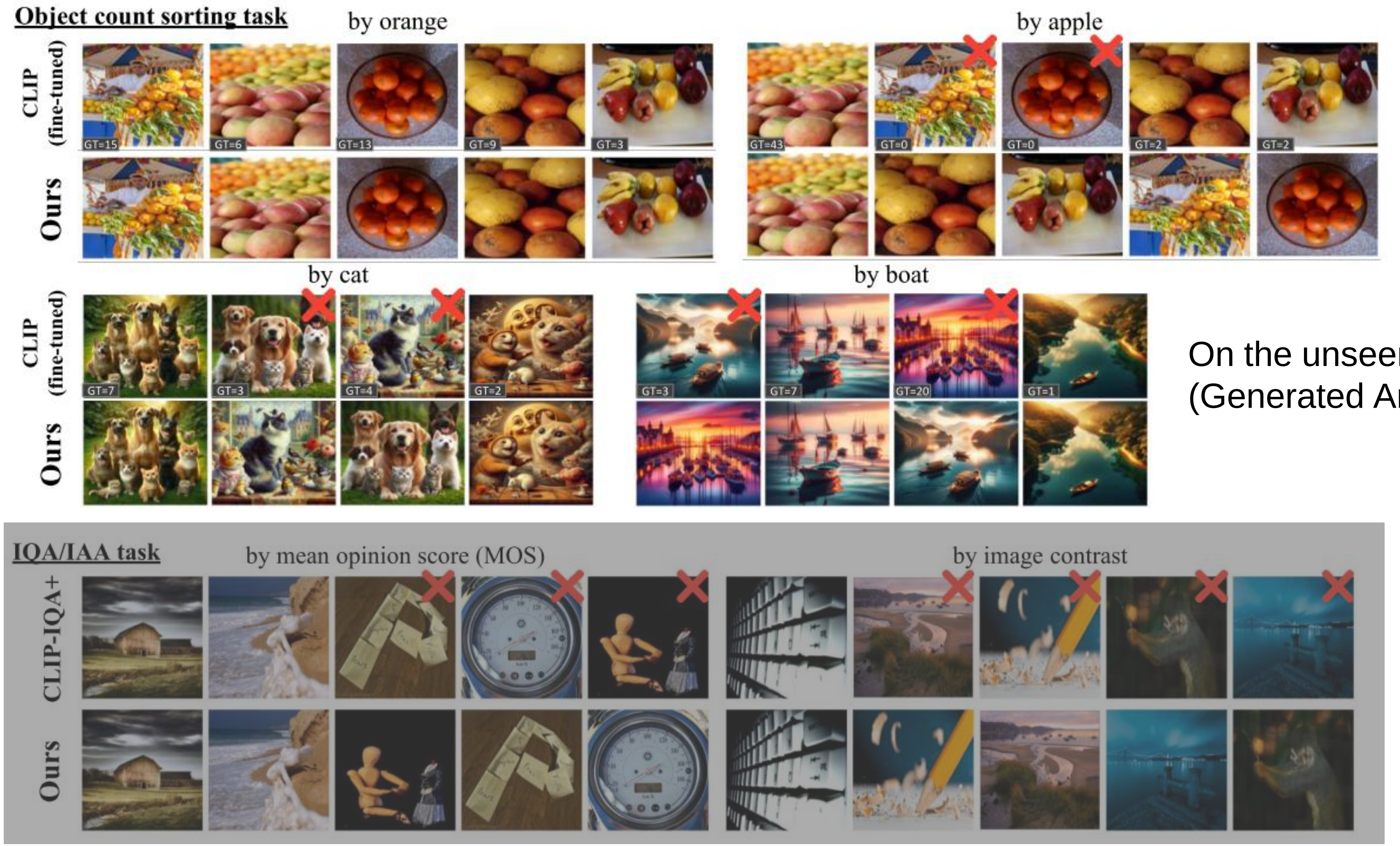
# Quantitative Results

| Method | Fine-tuning | PLCC (↑) | SRCC (↑) |
|---|---|---|---|
| BLIP-2 | | 0.284 | 0.252 |
| Flamingo (10-shot) | | 0.033 | 0.031 |
| InstructBLIP | | 0.509 | 0.485 |
| VLM-VILA | | 0.558 | 0.507 |
| Zero-shot CLIP | | 0.026 | 0.001 |
| CountingCLIP Paiss et al. (2023) | ✓ | 0.251 | 0.422 |
| Ours | ✓ | **0.624** | **0.557** |

| Method | Adience | | HCI | |
|---|---|---|---|---|
| | Accuracy (%) | MAE | Accuracy (%) | MAE |
| Zero-shot CLIP | 43.3 (3.6) | 0.80 (0.02) | 26.1 (0.6) | 1.48 (0.03) |
| CoOp | 60.6 (5.5) | 0.50 (0.08) | 51.9 (2.6) | 0.76 (0.06) |
| OrdinalCLIP | 61.2 (4.2) | 0.47 (0.06) | 56.4 (1.7) | 0.67 (0.03) |
| L2RCLIP | **66.2 (4.4)** | **0.36 (0.05)** | 67.2 (1.6) | 0.43 (0.03) |
| NumCLIP | - | - | 69.6 (2.0) | 0.35 (0.03) |
| InstructBLIP | 63.7 | 0.41 | 30.9 | 0.96 |
| Ours | 65.2 (2.9) | **0.36 (0.03)** | **72.8 (2.6)** | **0.32 (0.03)** |

| Method | Task-related pertaining | Fine-tuning | KonIQ-10k | | AVA Dataset | |
|---|---|---|---|---|---|---|
| | | | PLCC (↑) | SRCC (↑) | PLCC (↑) | SRCC (↑) |
| **Purely vision-based (task-specific)** | | | | | | |
| MUSIQ Ke et al. (2021) | | ✓ | 0.924 | 0.937 | 0.726 | 0.738 |
| **VLM-based (task-specific)** | | | | | | |
| VILA-P Ke et al. (2023) | ✓ | | - | - | 0.657 | 0.663 |
| VILA-R Ke et al. (2023) | ✓ | ✓ | **0.919** | **0.932** | **0.774** | **0.774** |
| CLIP (fine-tuned) | | ✓ | 0.245 | 0.216 | 0.162 | 0.160 |
| InstructBLIP Dai et al. (2023) | | | 0.211 | 0.163 | 0.229 | 0.226 |
| CLIP-IQA Wang et al. (2023c) | | | 0.695 | 0.727 | 0.420 | 0.415 |
| CLIP-IQA+ Wang et al. (2023c) | | ✓ | 0.895 | 0.909 | 0.677 | 0.587 |
| Hentschel et al. (2022) | | ✓ | - | - | 0.731 | 0.741 |
| Ours | | ✓ | **0.919** | **0.911** | **0.760** | **0.747** |

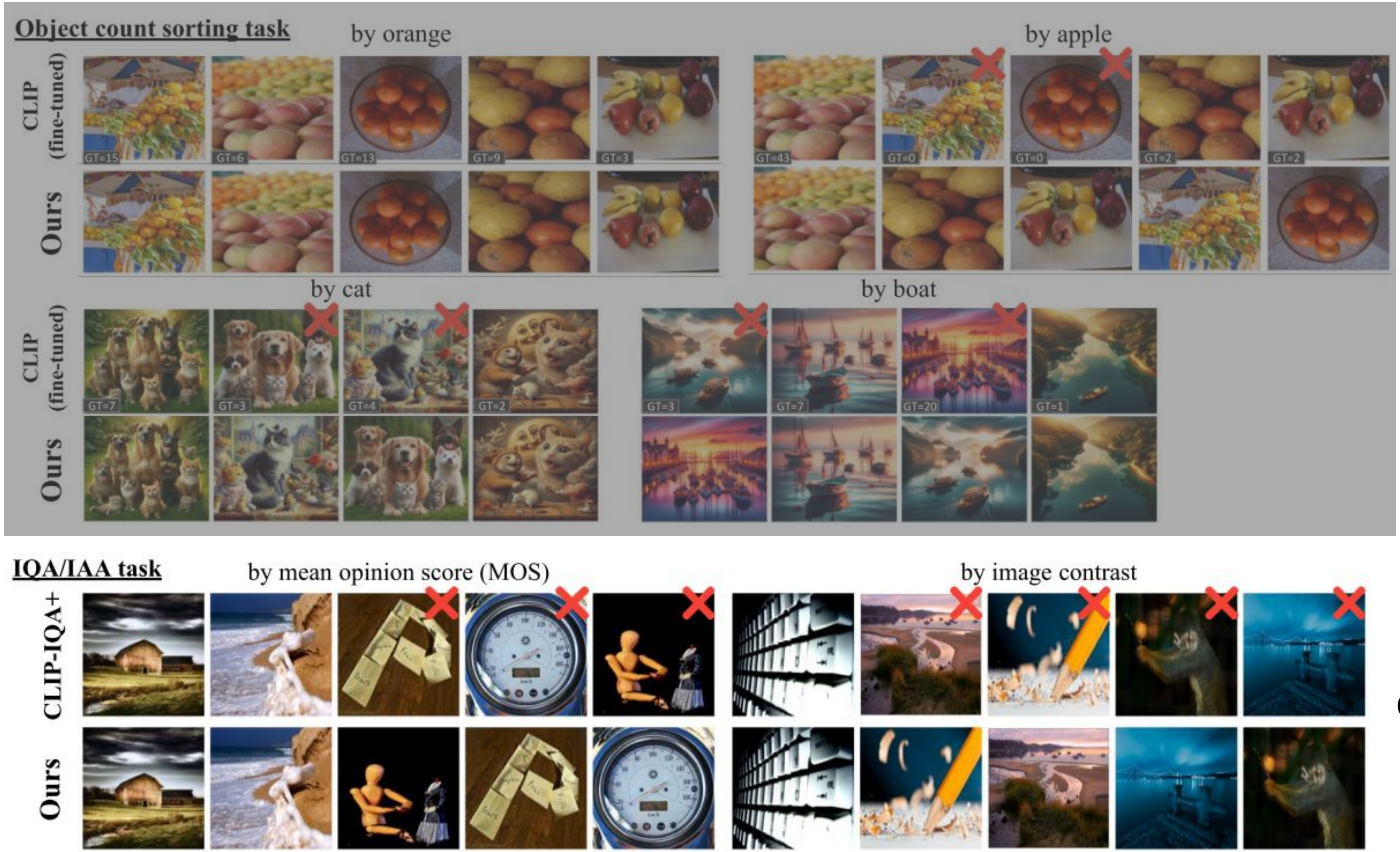
# Qualitative Results




Same set of images w/ different query

On the unseen domain (Generated Arts)

# Qualitative Results

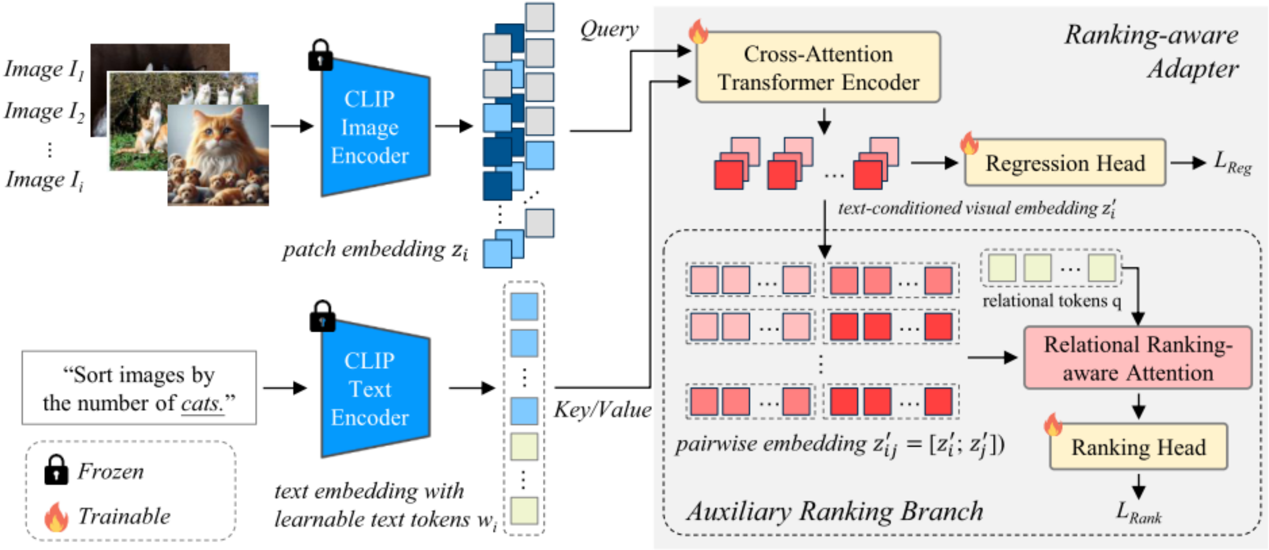



On IQA Tasks

# Ablation Study

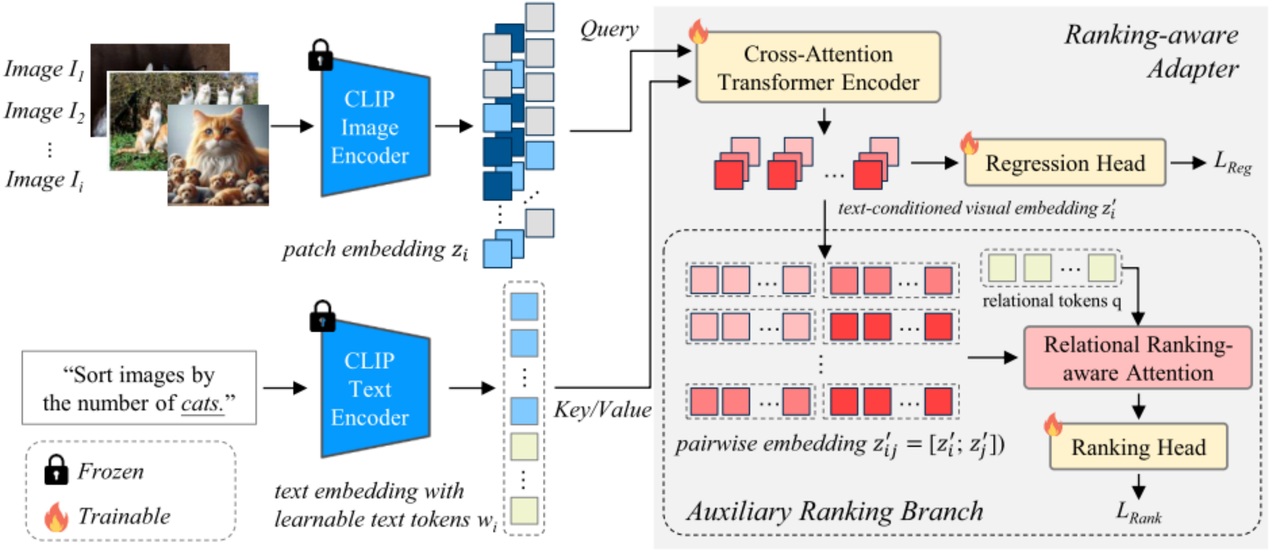
| LTR paradigm | Ranking Head | Ranking-aware Attention | HCI MAE (↓) | Object count sorting PLCC (↑) | Object count sorting SRCC (↑) |
|---|---|---|---|---|---|
| - | - | - | 1.113 | 0.251 | 0.422 |
| ✓ | - | - | 0.402 | 0.612 | 0.538 |
| ✓ | ✓ | - | 0.355 (+11.69%) | 0.619 (+1.14%) | 0.536 (−0.37%) |
| ✓ | ✓ | ✓ | 0.317 (+21.14%) | 0.624 (+1.96%) | 0.557 (+3.53%) |

Baseline: CLIP w/ finetuning (contrastive Learning)



# Ablation Study

| | | | HCI | | Object count sorting | | | |
|---|---|---|---|---|---|---|---|---|
| LTR paradigm | Ranking Head | Ranking-aware Attention | MAE (↓) | | PLCC (↑) | | SRCC (↑) | |
| - | - | - | 1.113 | | 0.251 | | 0.422 | |
| ✓ | - | - | 0.402 | | 0.612 | | 0.538 | |
| ✓ | ✓ | - | 0.355 | (+11.69%) | 0.619 | (+1.14%) | 0.536 | (−0.37%) |
| ✓ | ✓ | ✓ | 0.317 | (+21.14%) | 0.624 | (+1.96%) | 0.557 | (+3.53%) |



# Conclusion

- We present an efficient and scalable framework **for text-driven image ranking** by reframing CLIP’s image-text contrastive learning into an **LTR task**.
- By leveraging a lightweight adapter with our **ranking-aware attention** module, it can effectively capture **text-driven visual differences between image pairs**.
- Our work, **an all-in-one and end-to-end method**, surpass CLIP baselines and achieve results comparable to **SOTA methods tailored for specific task**, highlights the potential of leveraging VLMs with visual distinctions for developing sense of number sense.

# Thank You
# For Your Attention!

Paper

Code