TRACE: Text-Region Alignment with Conceptual Explainability

Duy A. Nguyen, Huyen Nguyen
Instructor: Prof. Minh Do

University of Illinois Urbana-Champaign

May 6, 2025

Text Grounding Task

Text Grounding: Identify image regions that correspond to a given text token or keyword.

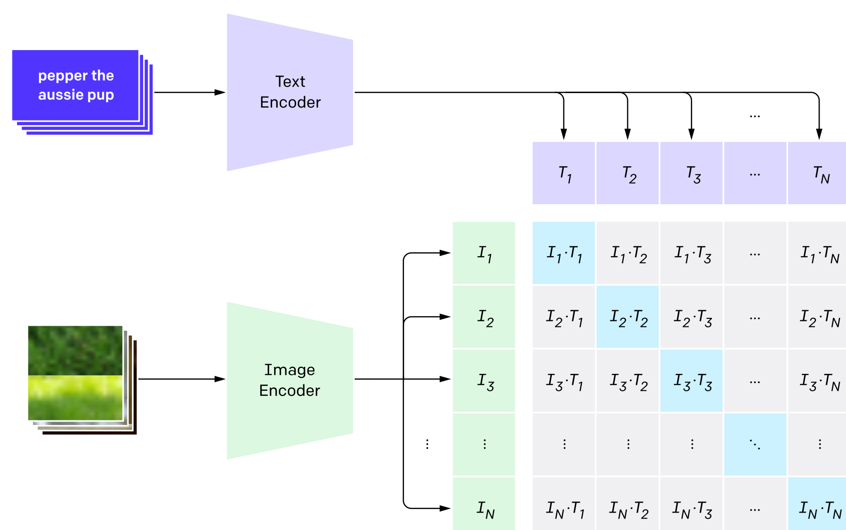
Applications:

- **Search and retrieval:** Highlight relevant areas based on natural language queries.
- **Robotics and AR:** Enable spatial understanding from verbal commands.
- **Medical imaging:** Localize findings from text-based reports.

Text Grounding Paradigms

Setting	Close-vocab	Open-vocab
Image–Caption Grounding		
Patch–Token Grounding	✓	

1. Contrastive pre-training



Transparent and Controllable Text Grounding

The Challenge: Most deep learning models (e.g., CLIP) perform grounding as an *emergent behavior* of large-scale training — yet offer little insight into **why** a word is grounded to a region.

- Deep models like CLIP can align words and regions — but they are **black-box**.
- They reveal **where** attention goes, but not **what** it means or **how** it's structured.

We ask: Can grounding be made transparent and controllable — not just accurate?

- **Interpretability**: Understandable rationale behind predictions.
- **Auditability**: Clear trace from token to region.
- **Controllability**: Ability to steer model behavior.

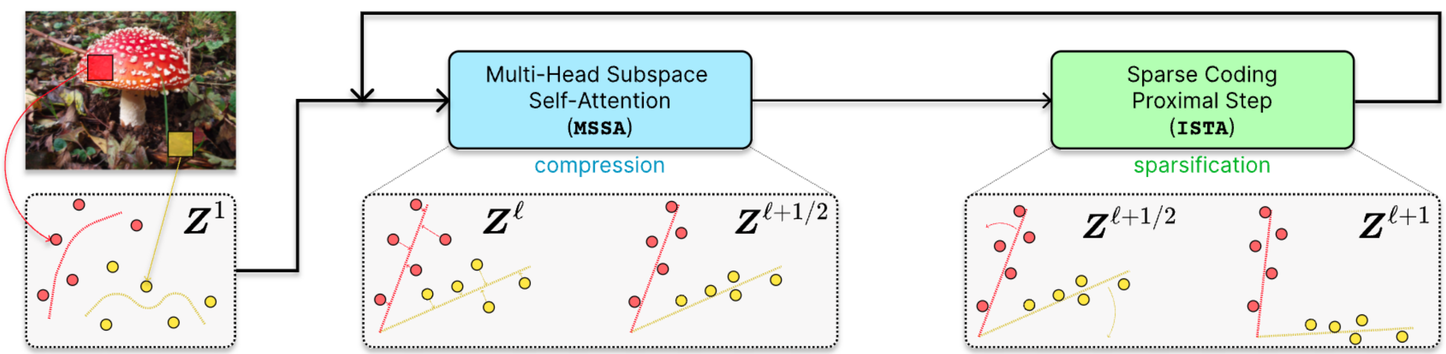
HOW?

CRATE: A White-Box Vision Transformer [4]

- **Goal:** Build a fully interpretable architecture from first principle for representation learning.
- **Key Idea:** CRATE optimizes a **sparse rate reduction** objective, compressing tokens into a sparse combination of K low-dimensional subspaces defined by orthonormal bases $\mathbf{U}_{[K]} = \{\mathbf{U}_1, \ldots, \mathbf{U}_K\}$:

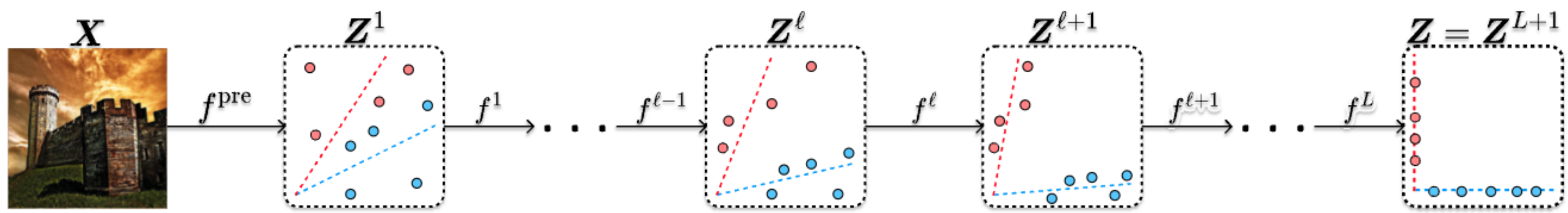
$$\max_f O_1(\mathbf{X}) = \mathbb{E}_{\mathbf{Z}=f(\mathbf{X})}\left[R(\mathbf{Z}) - R^c(\mathbf{Z} \mid \mathbf{U}_{[K]}) - \lambda\|\mathbf{Z}\|_0\right] \quad (1)$$

  - $R(\mathbf{Z})$: coding rate of uncompressed representation.
  - $R^c(\mathbf{Z} \mid \mathbf{U}_{[K]})$: rate after projecting onto learned subspaces.
  - $\|\mathbf{Z}\|_0$: sparsity penalty to encourage minimal activation.



Conceptual illustration of CRATE

CRATE Alternated Optimization Procedure



CRATE rollout illustration

- **Forward pass:** $\boldsymbol{X} \xrightarrow{f^0} \boldsymbol{Z}^0 \to \cdots \to \boldsymbol{Z}^\ell \xrightarrow{f^\ell} \boldsymbol{Z}^{\ell+1} \to \cdots \to \boldsymbol{Z}^L = \boldsymbol{Z}$
  - Each layer performs an approximation for a single gradient backpropagation step to optimize Objective 1: $\boldsymbol{Z}^{\ell+1} = f^\ell(Z^\ell) \approx \boldsymbol{Z}^\ell - \kappa \nabla_{\boldsymbol{Z}} O_1 \left(\boldsymbol{Z}^\ell \right)$
  - Last output $\boldsymbol{Z}$ is most structured and disentangled features: $\boldsymbol{Z} = \sum_k^K \alpha_k \mathbf{U}_k \mathbf{U}_k^T \boldsymbol{Z}^{L-1}$, where $\alpha = [\alpha_1, \ldots, \alpha_K]^T \approx [0, \ldots, 1, \ldots, 0]^T$
- **Backward pass:** Learn $\mathbf{U}_{[K]}$ via ordinary data-driven gradient-based approach. e.g. CE loss for classification task: $L_{\mathrm{CE}}\left(f, f^{\text{head}}\right) \doteq \mathbb{E}_{\boldsymbol{X},\boldsymbol{y}}\left[H\left(\boldsymbol{y}, \text{softmax}\left\{\left(f^{\text{head}} \circ f\right)(\boldsymbol{X})\right\}\right)\right]$.

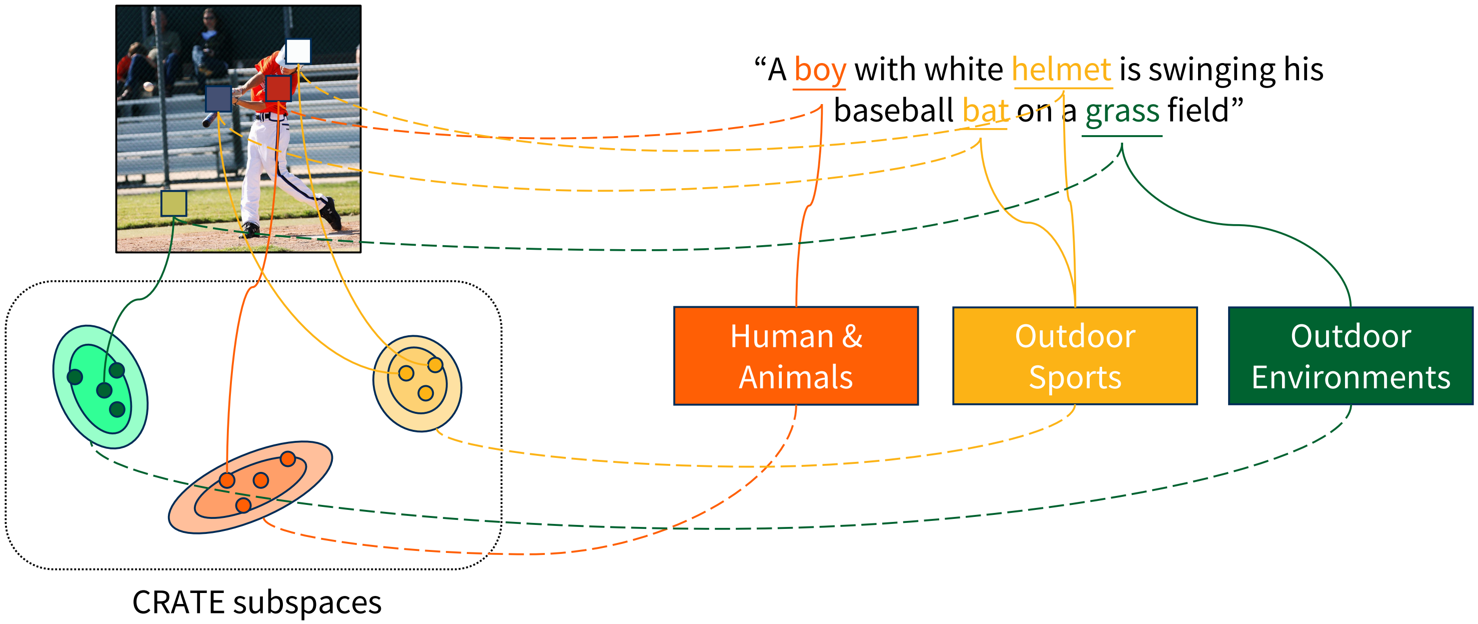
Motivating Question

- CRATE offers the structural transparency we need – it shows **how** patches are encoded via low-dimensional subspaces.
- However, it does not explain **what** those subspaces represent — they remain abstract and data-driven.
- It lacks **semantic interpretability**: no clear mapping between subspaces and human concepts.
- **Can we bring semantic meaning into CRATE's structure** for grounding task — while preserving its mathematical transparency?

Language as Semantic Bridge

- Language naturally encodes **human-consumable semantics**—structured, interpretable, and compositional.
- Can we use language as a **semantic bridge** to bring interpretable meaning to CRATE's visual subspaces?

Note: Linguistic concepts are inherently hierarchical (e.g., $\mathcal{vehicle} \rightarrow$ $\mathcal{car}$, $\mathcal{bus}$)—a structure that mirrors CRATE's subspace decomposition.



Our Contribution: Multi-Resolution Semantic Alignment

Core Idea:

- **Cluster-level alignment:** Assign CRATE's visual subspaces to *coarse-grained* concepts (e.g., `vehicle`, `animal`).
- **Token-level alignment:** Align individual patch embeddings to *fine-grained* child concepts (e.g., `car`, `dog`) within their superclass.

Scope and Assumptions:

- Operate under **weak supervision**: only image-caption pairs are required.
- Leverage **frozen text embeddings** (e.g., GloVe or Word2Vec) to serve as stable semantic anchors during training.

The resulting framework is named *TRACE (Text-Region Alignment with Conceptual Explainability)*.

Preliminary Result: Quantitative Evaluation

Quantitative result: w/ Threshold = 0.5 (binary mask)

Method	**Precision**	**Recall**	**mIoU**
CLIP	10.50	55.84	9.39
TRACE	**14.59**	**61.15**	**12.94**

Observations:

- TRACE improves both precision (more selective) and recall (more complete).
- High recall but low precision/mIoU in both methods suggests over-prediction:
  - Large/multiple regions are assigned per token.
  - Many false positives inflate recall, but hurt precision.

Conclusion

- Introduced semantically-grounded CRATE extension, named TRACE.
- Preserves structure and adds interpretability.
- TRACE show potential in fine-grain Text-Grounding task, with weakly supervised signal.
- Future work:
    - Further loss enhancement, making cluster more separated.
    - Tackling multi-resolution alignment implicitly.
    - Extend to open-vocab text-grounding.

References I

[1] Ioana Bica et al. "Improving fine-grained understanding in image-text pre-training". In: *Forty-first International Conference on Machine Learning*.

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. "Coco-stuff: Thing and stuff classes in context". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1209–1218.

[3] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[4] Yaodong Yu et al. "White-box transformers via sparse rate reduction". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 9422–9457.