VISTA: Visual-Textual Knowledge Graph Representation Learning

Jaejun Lee, **Chanyoung Chung**, **Hochang Lee**, **Sungho Jo**, and **Joyce Jiyoung Whang***

▼ BDILab ▼ GitHub

* Corresponding Author

School of Computing, KAIST

The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)

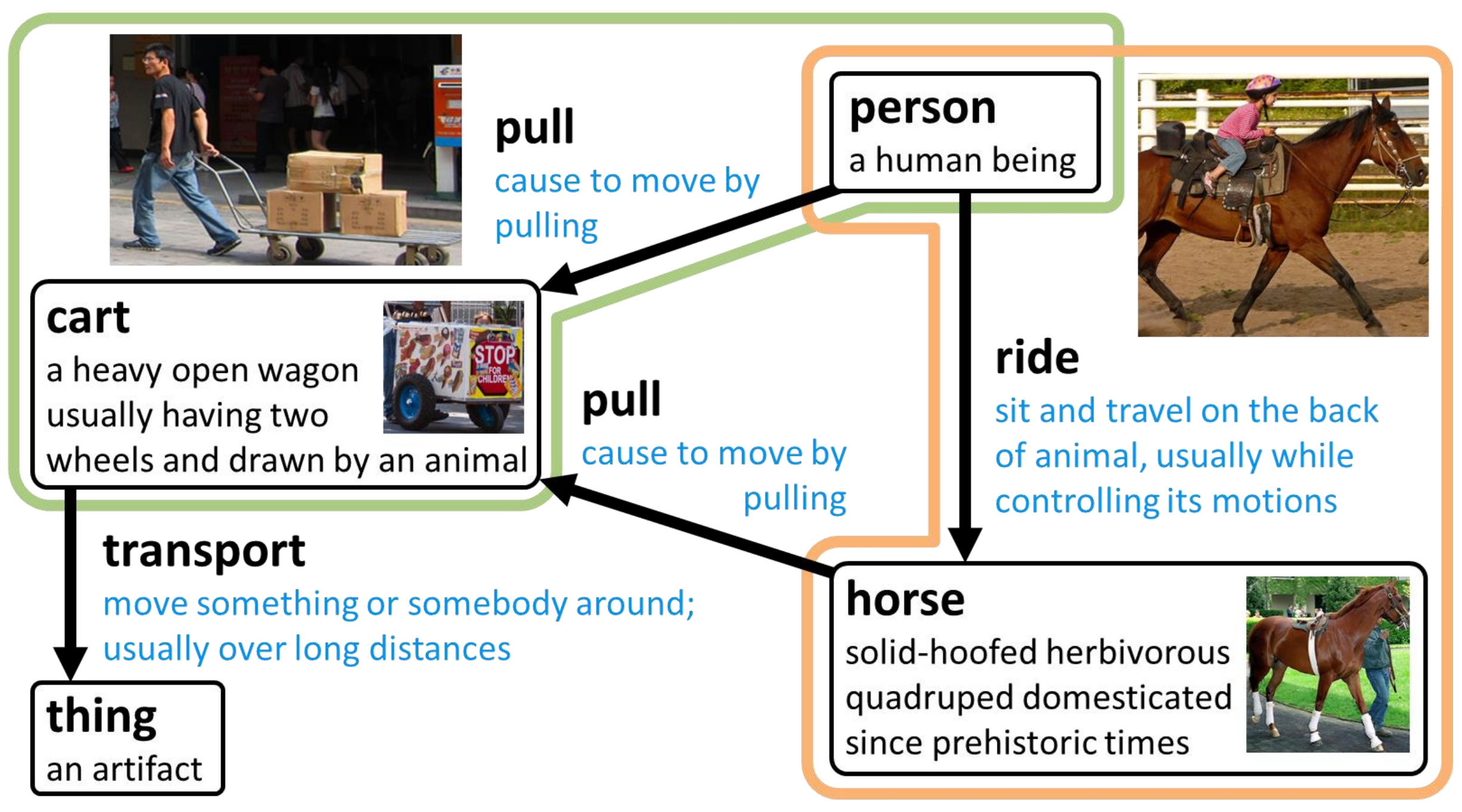
KAIST BDILab BIG DATA INTELLIGENCE

Main Contributions

- Define **Visual-Textual Knowledge Graphs** (**VTKGs**)
  - Create two real-world VTKG datasets: **VTKG-I** and **VTKG-C**
- Propose **VIS**ual-**T**extu**A**l (**VISTA**) knowledge graph representation learning method that utilizes **visual and textual features of relations and entities**.
  - Define entity encoding transformer, relation encoding transformer, and triplet decoding transformer to predict a missing entity in a triplet.
- VISTA outperforms **10 different** state-of-the-art knowledge graph completion methods, including multimodal knowledge graph embedding methods.

Visual-Textual Knowledge Graphs

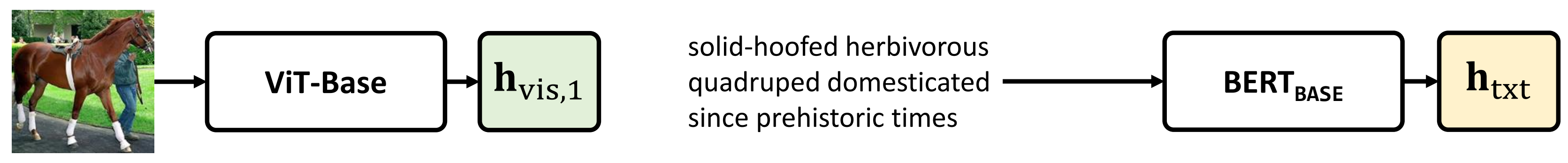
- **Visual-Textual Knowledge Graphs** (**VTKGs**)
  - Entities and triplets in a VTKG can be represented by **images**.
  - Entities and relations have their **text descriptions**.


pull
cause to move by pulling
person
a human being
cart
a heavy open wagon usually having two wheels and drawn by an animal
pull
cause to move by pulling
ride
sit and travel on the back of animal, usually while controlling its motions
transport
move something or somebody around; usually over long distances
horse
solid-hoofed herbivorous quadruped domesticated since prehistoric times
thing
an artifact

- **Link Prediction on VTKGs:** Predicting missing links between entities
  - e.g., Given an incomplete triplet ⟨horse, pull, ?⟩, predict ? as "thing"
- **Creating Real-World VTKGs**
  - Extract **visual commonsense knowledge** using four different computer vision benchmark datasets: **VRD**, **UnRel**, **HICO-DET**, **VisKE**
  - Add triplets from **WordNet** and **ConceptNet**

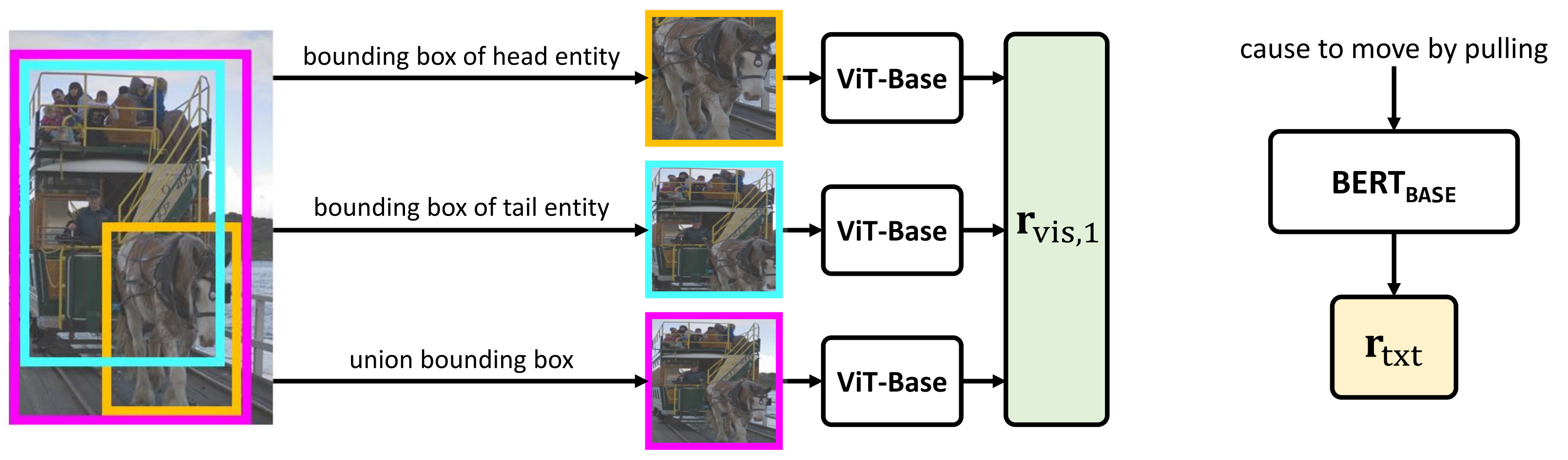
Extracting Visual and Textual Features

- Extracting visual and textual features of an entity


ViT-Base
$\mathbf{h}_{\text{vis},1}$
solid-hoofed herbivorous quadruped domesticated since prehistoric times
$\text{BERT}_{\text{BASE}}$
$\mathbf{h}_{\text{txt}}$

visual features of "horse" textual feature of "horse"

- Extracting visual and textual features of a relation

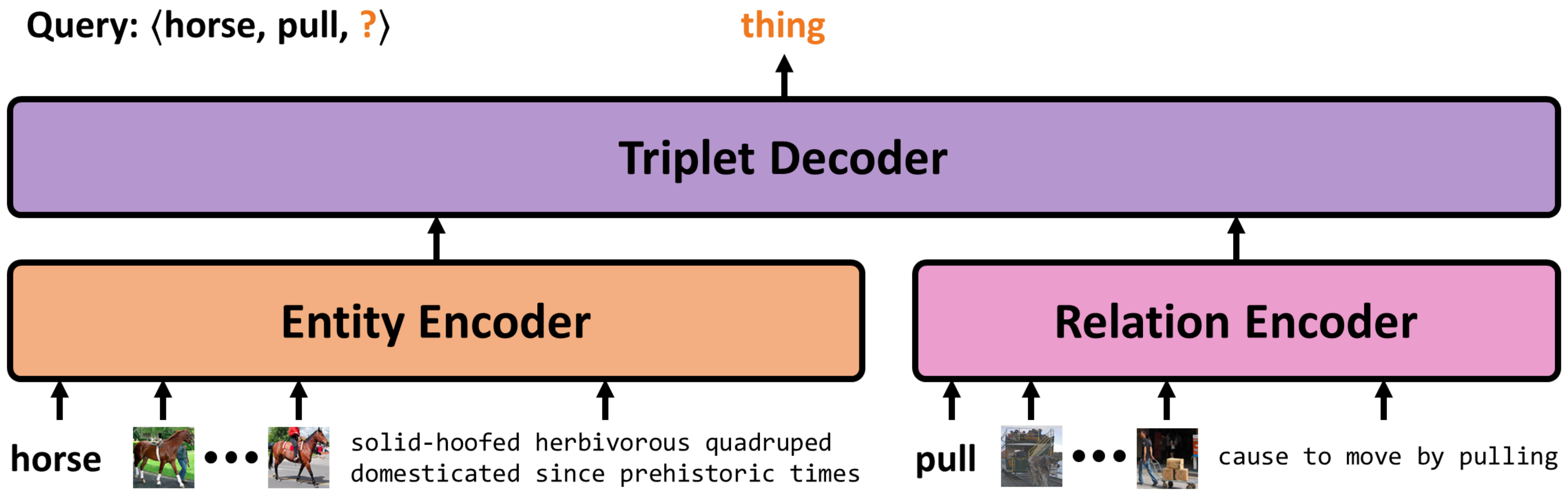

bounding box of head entity
bounding box of tail entity
union bounding box
ViT-Base
ViT-Base
ViT-Base
$\mathbf{r}_{\text{vis},1}$
cause to move by pulling
$\text{BERT}_{\text{BASE}}$
$\mathbf{r}_{\text{txt}}$

visual features of "pull" textual feature of "pull"

Overview of VISTA

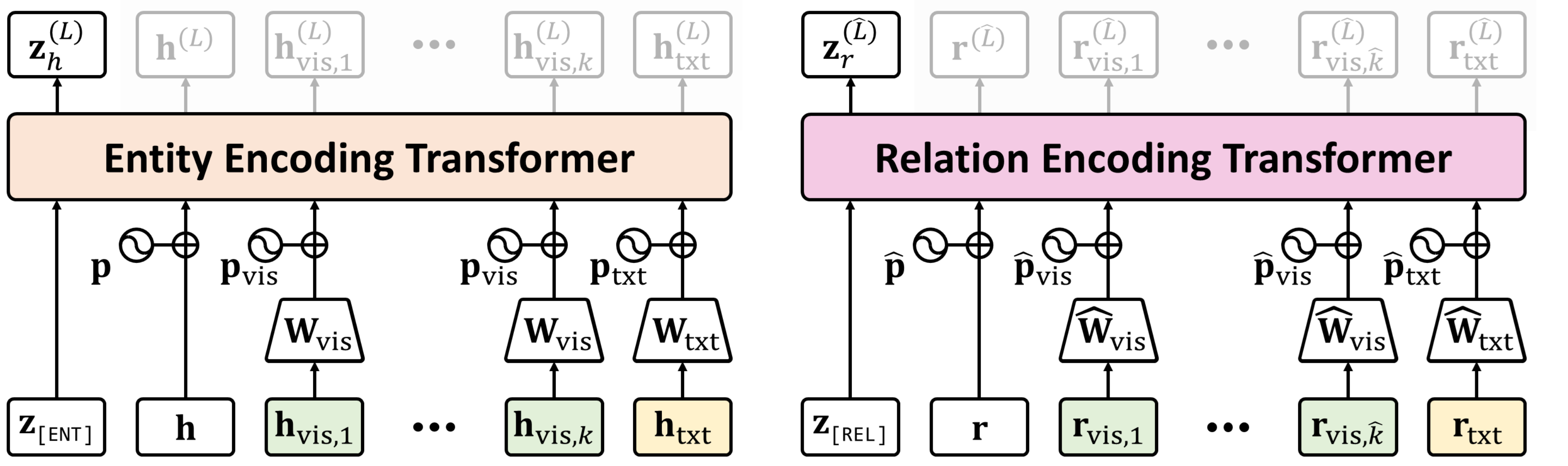
- **Entity/Relation Encoder**
  - Calculate the representations of entities and relations by **an entity encoding transformer** and **a relation encoding transformer**
- **Triplet Decoder**
  - Predict a missing entity in a triplet using **a triplet decoding transformer**

Query: ⟨**horse, pull, ?**⟩



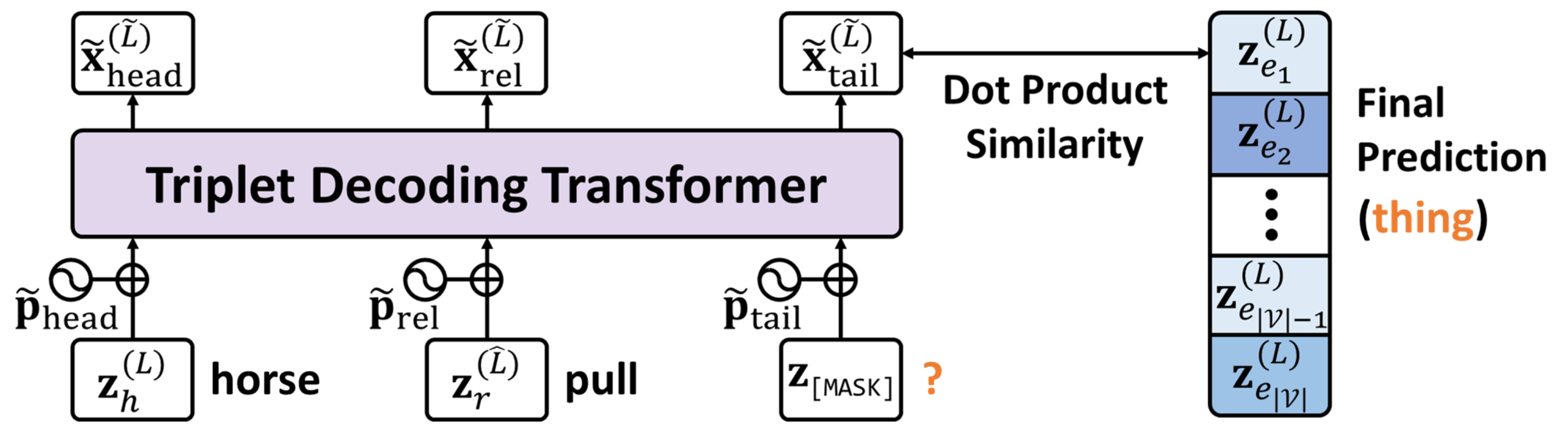
Entity/Relation Encoder

- **Entity (Relation) Encoding Transformer**
  - Compute the representation of an input entity (relation) by considering its learnable vector, visual features, and textual feature.


$\mathbf{z}_h^{(L)}$
$\mathbf{h}^{(L)}$
$\mathbf{h}_{\text{vis},1}^{(L)}$
$\mathbf{h}_{\text{vis},k}^{(L)}$
$\mathbf{h}_{\text{txt}}^{(L)}$
Entity Encoding Transformer
$\mathbf{p}$
$\mathbf{p}_{\text{vis}}$
$\mathbf{p}_{\text{vis}}$
$\mathbf{p}_{\text{txt}}$
$\mathbf{W}_{\text{vis}}$
$\mathbf{W}_{\text{vis}}$
$\mathbf{W}_{\text{txt}}$
$\mathbf{z}_{[\text{ENT}]}$
$\mathbf{h}$
$\mathbf{h}_{\text{vis},1}$
$\mathbf{h}_{\text{vis},k}$
$\mathbf{h}_{\text{txt}}$
$\mathbf{z}_r^{(L)}$
$\mathbf{r}^{(L)}$
$\mathbf{r}_{\text{vis},1}^{(L)}$
$\mathbf{r}_{\text{vis},\hat{k}}^{(L)}$
$\mathbf{r}_{\text{txt}}^{(L)}$
Relation Encoding Transformer
$\hat{\mathbf{p}}$
$\hat{\mathbf{p}}_{\text{vis}}$
$\hat{\mathbf{p}}_{\text{vis}}$
$\hat{\mathbf{p}}_{\text{txt}}$
$\widehat{\mathbf{W}}_{\text{vis}}$
$\widehat{\mathbf{W}}_{\text{vis}}$
$\widehat{\mathbf{W}}_{\text{txt}}$
$\mathbf{z}_{[\text{REL}]}$
$\mathbf{r}$
$\mathbf{r}_{\text{vis},1}$
$\mathbf{r}_{\text{vis},\hat{k}}$
$\mathbf{r}_{\text{txt}}$

Triplet Decoder

- **Triplet Decoding Transformer**
  - Predict a missing entity based on the entity/relation representations


$\widetilde{\mathbf{x}}_{\text{head}}^{(\tilde{L})}$
$\widetilde{\mathbf{x}}_{\text{rel}}^{(\tilde{L})}$
$\widetilde{\mathbf{x}}_{\text{tail}}^{(\tilde{L})}$
Dot Product Similarity
Triplet Decoding Transformer
$\widetilde{\mathbf{p}}_{\text{head}}$
$\widetilde{\mathbf{p}}_{\text{rel}}$
$\widetilde{\mathbf{p}}_{\text{tail}}$
$\mathbf{z}_h^{(L)}$ horse
$\mathbf{z}_r^{(L)}$ pull
$\mathbf{z}_{[\text{MASK}]}$?
$\mathbf{z}_{e_1}^{(L)}$
$\mathbf{z}_{e_2}^{(L)}$
$\mathbf{z}_{e_{|\mathcal{V}|-1}}^{(L)}$
$\mathbf{z}_{e_{|\mathcal{V}|}}^{(L)}$
Final Prediction (thing)

Experiments

- **Baseline methods**: ANALOGY, ComplEx-N3, RotatE, PairRE, RSME, TransAE, MKGformer, OTKGE, MoSE, IMF
- **Knowledge Graph Completion on VTKGs**

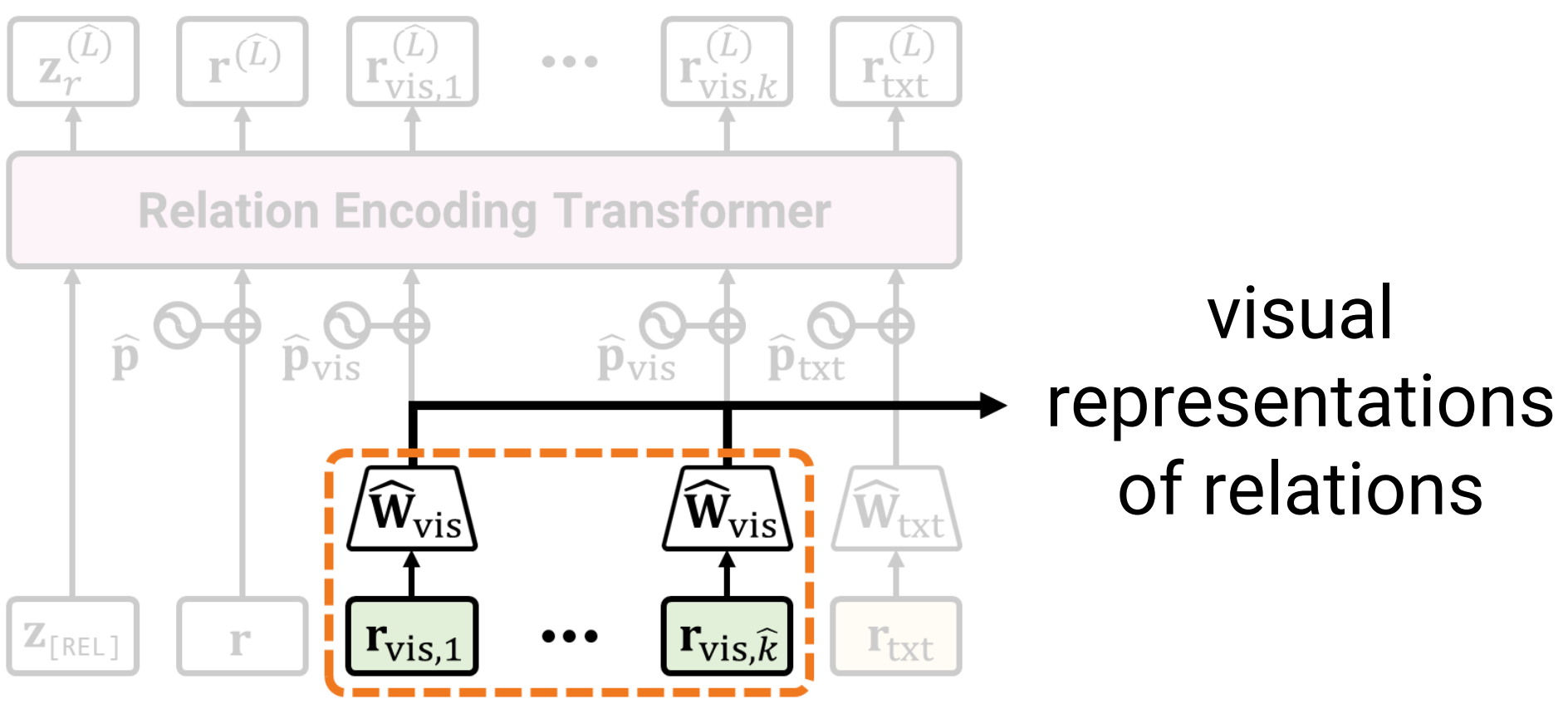
		MRR (↑)	Hit@10 (↑)	Hit@3 (↑)	Hit@1 (↑)	MR (↓)
VTKG-I	Best-baseline	0.4306	0.3588	0.4656	0.6374	19.5
	VISTA	**0.4650**	**0.3626**	**0.5076**	**0.6641**	**17.3**
VTKG-C	Best-baseline	0.4227	0.3706	0.4762	0.5977	527.0
	VISTA	**0.4675**	**0.3918**	**0.4961**	**0.6157**	**220.8**

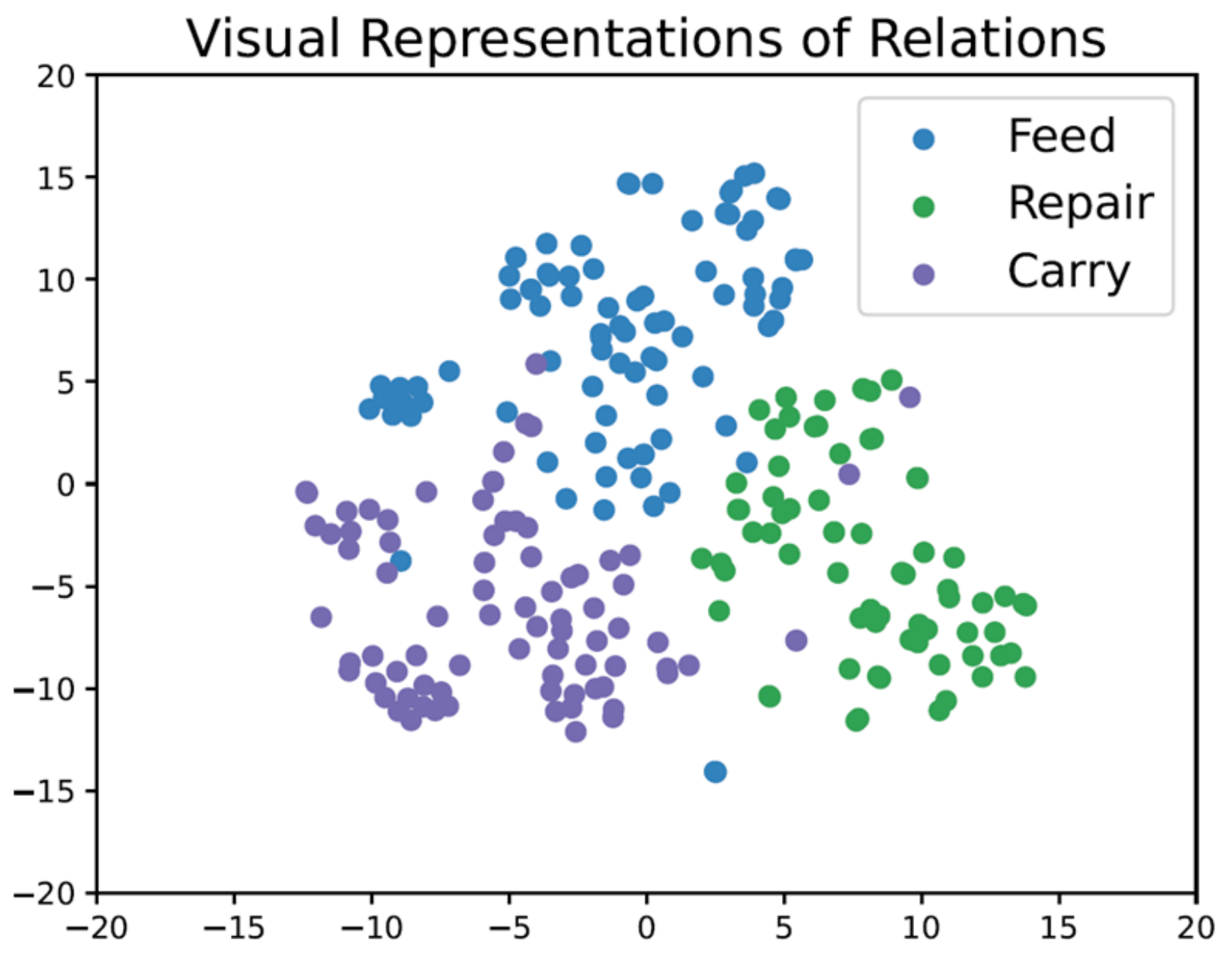
- **Knowledge Graph Completion on Existing Benchmark Datasets**

		MRR (↑)	Hit@10 (↑)	Hit@3 (↑)	Hit@1 (↑)	MR (↓)
WN18RR++	Best-baseline	0.5308	0.4697	0.5557	0.6681	**108.0**
	VISTA	**0.5526**	**0.4871**	**0.5799**	**0.6755**	177.6
FB15K237	Best-baseline	0.3677	0.2735	0.4040	0.5573	132.3
	VISTA	**0.3808**	**0.2873**	**0.4158**	**0.5718**	**114.2**

Qualitative Analysis

- **Visual Representation Vectors of Relations**
  - Visual representation vectors of relations are **well-clustered**.


Relation Encoding Transformer
$\mathbf{r}_{\text{vis},1}$
$\mathbf{r}_{\text{vis},\hat{k}}$
visual representations of relations



- **Top Similar Entities/Relations**
  - VISTA returns the most semantically close entities and relations.

Query		BERT	ViT	VISTA
	1	incense	leisure_wear	orange
dark_red	2	coloring	sportswear	red
	3	buffer	sweatshirt	crimson

top similar entities

Query		BERT	ViT	VISTA
	1	move	straddle	keep
have	2	influence	hop_on	hold
	3	begin	inspect	incorporate

top similar relations

Conclusion

- Introduce **Visual-Textual Knowledge Graphs** (**VTKGs**).
- Propose **VIS**ual-**T**extu**A**l (**VISTA**) knowledge graph representation learning method to solve knowledge graph completion problems in **real-world VTKGs**.
- VISTA takes into account **visual and textual features** of entities and relations.
- VISTA substantially outperforms **10 different** state-of-the-art methods.

GitHub: https://github.com/bdi-lab/VISTA | Lab Homepage: https://bdi-lab.kaist.ac.kr