Lecture#1: Knowledge Graph Embedding with Multimodal Data

Joyce Jiyoung Whang
School of Computing, KAIST

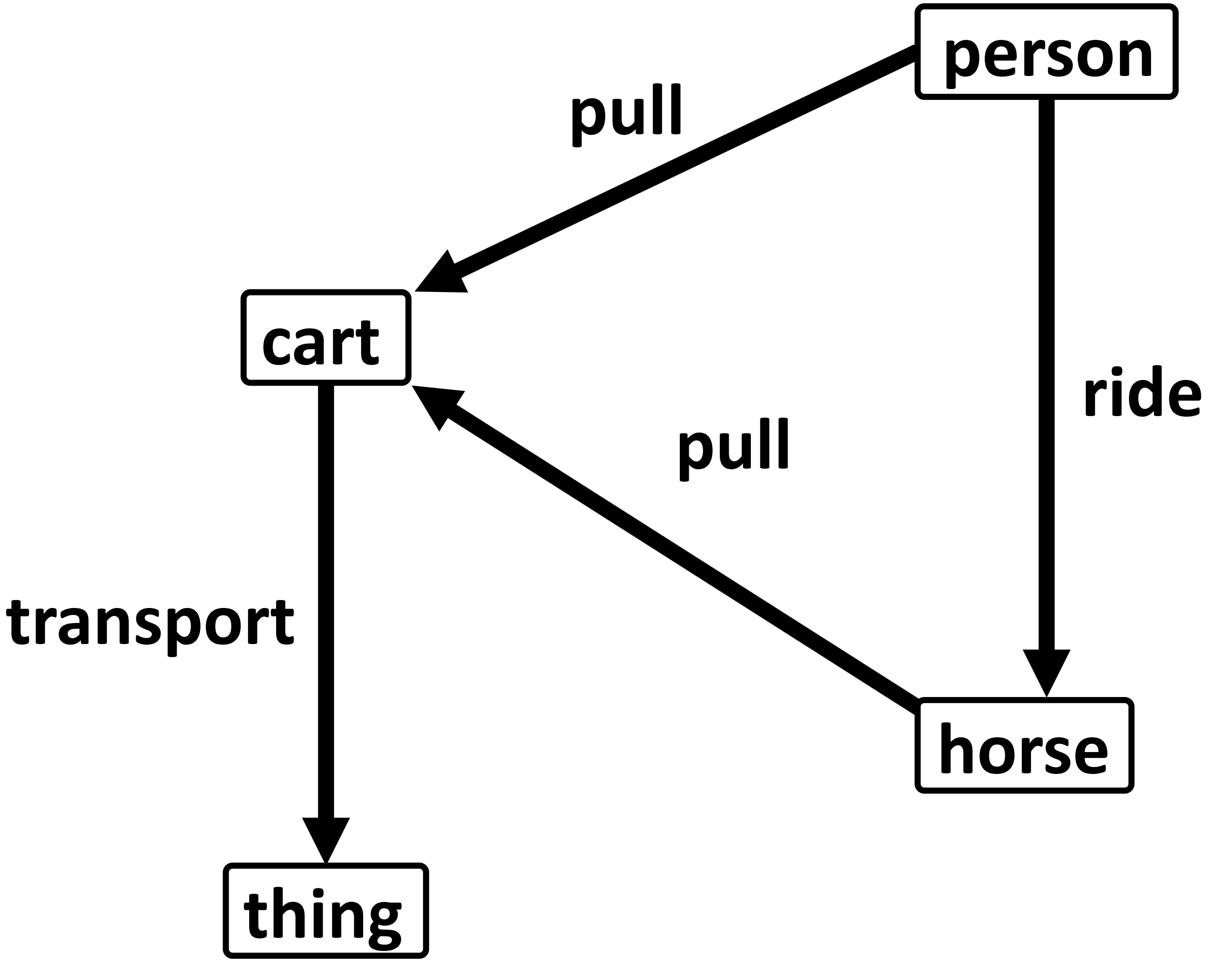
Key Facets in Modern Knowledge Graph Representation Learning (KeyKGRL), ISWC 2025 Tutorial

https://bdi-lab.kaist.ac.kr

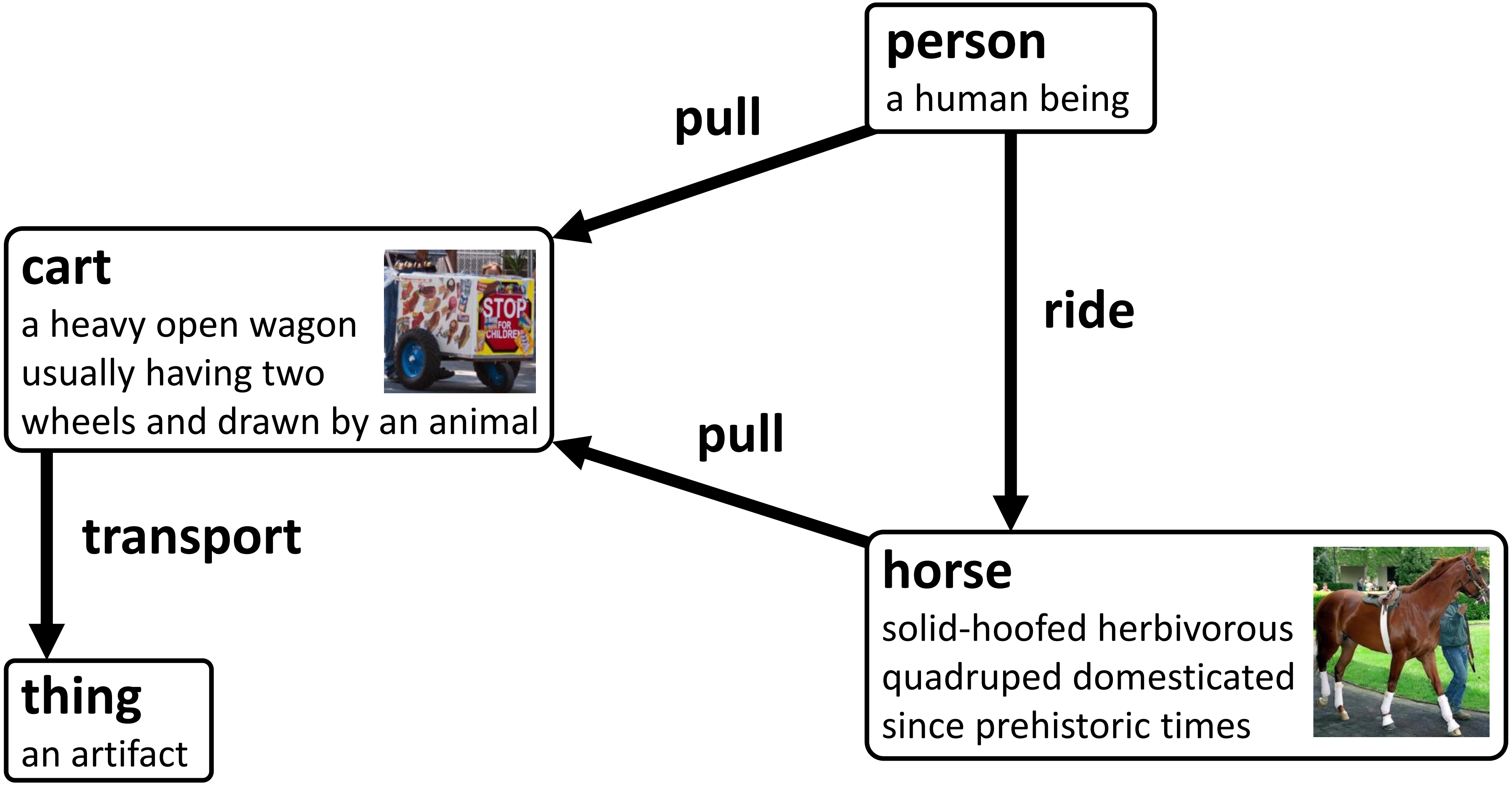
KAIST BDILab
BIG DATA INTELLIGENCE

01

Knowledge Graphs



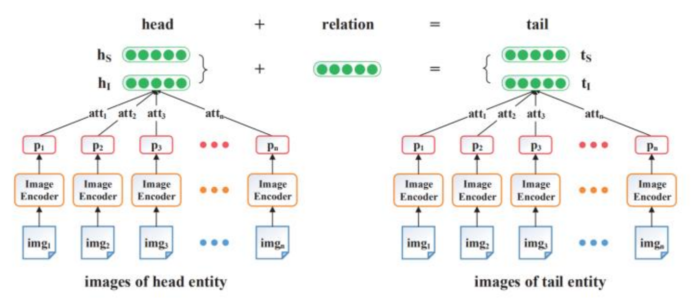
01 Multimodal Knowledge Graphs



01 Multimodal KG Embedding Methods

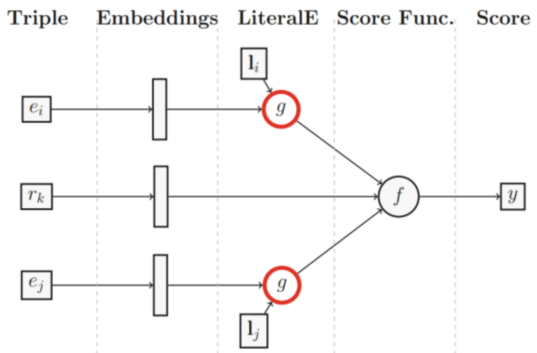
IKRL (IJCAI 2017)

- Each entity has visual embeddings and structural embeddings
- Computes visual embeddings of entities by attentively aggregating the visual features
- Utilizes both visual and structural embeddings of head and tail entities to compute plausibility of a triplet



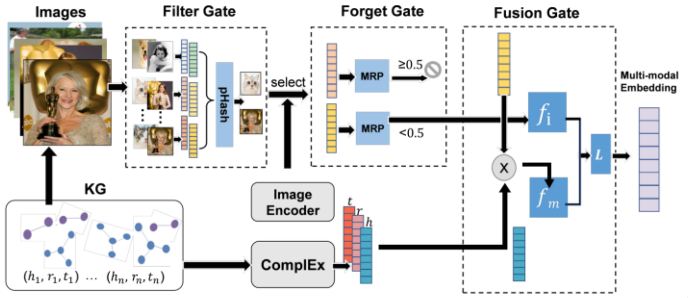
LiteralE (ISWC 2019)

- Randomly selects one modality instance per modality type
- Combines the entity embeddings with its modality vectors using a gating mechanism
- Assumes that all entities have some values for all modalities



RSME (MM 2021)

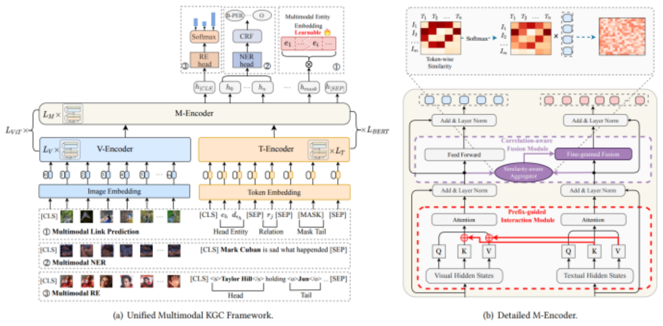
- Selects one image per entity based on the pairwise similarities
- Finds relations that benefit from the visual features, and does not use the visual features otherwise
- Considers the plausibility of the triplet and the similarity of visual features of the head and tail entities



01 Multimodal KG Embedding Methods

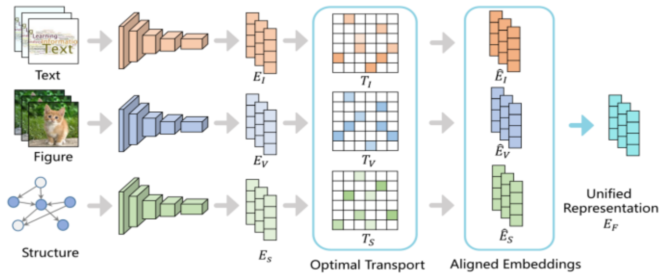
MKGformer (SIGIR 2022)

- Utilizes pretrained ViT and BERT while freezing their parameters
- Learns the embeddings of entities that align with the feature space of the pretrained BERT
- The latter layers of BERT also consider the vectors from ViT for cross-modal interaction



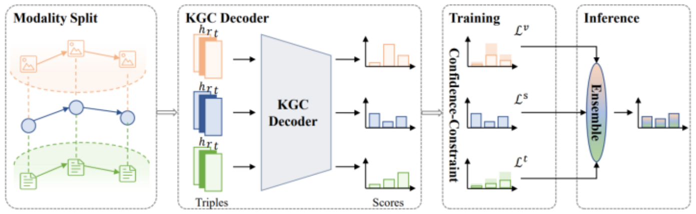
OTKGE (NeurIPS 2022)

- Models multimodal fusion procedure as a transportation plan
- Moves different modal embeddings to a unified space by minimizing the distance between modalities
- The distance minimization maintains consistency and comprehensiveness of modalities



MoSE (EMNLP 2022)

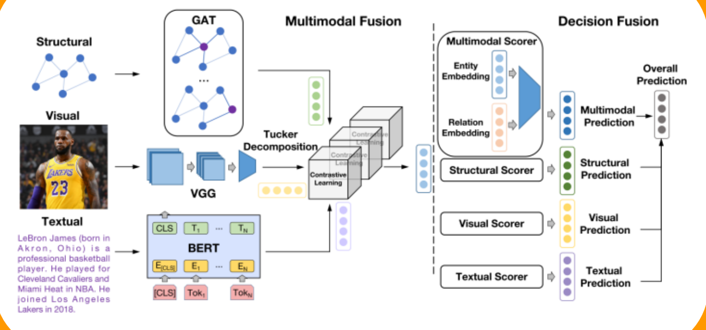
- Learns different relation embeddings for each modality to alleviate modality interference
- Makes per-modality predictions and exploits various ensemble methods to combine the predictions
- Models the modality importance dynamically



01 Multimodal KG Embedding Methods

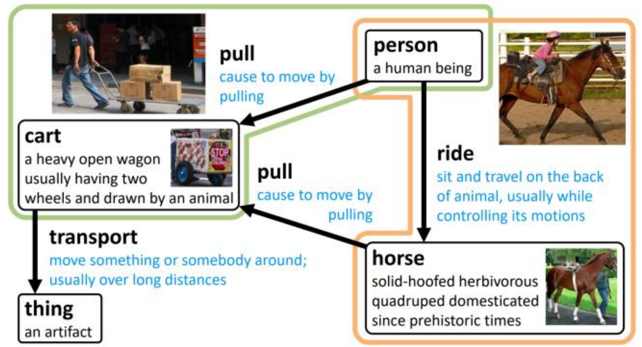
IMF (TheWebConf 2023)

- Two-stage multimodal fusion framework that preserves modality-specific knowledge
- Uses both the individual modalities and the fused representation
- Learns weights for each modality and takes weighted average over the predictions of all modalities



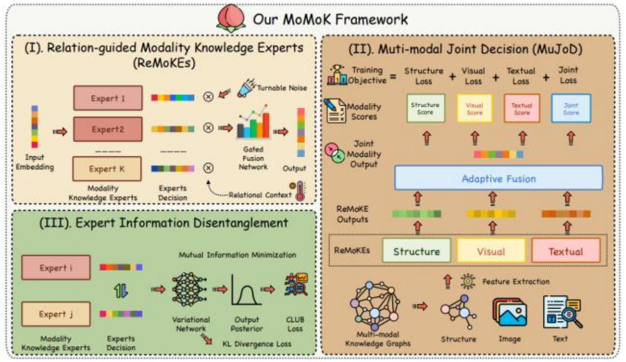
VISTA (EMNLP 2023 Findings)

- Proposes Visual-Textual Knowledge Graph (VTKG)
  - Not only entities, but also triplets can be explained using images
  - Both entities and relations accompany text descriptions
- Incorporates visual and textual features of entities and relations



MoMoK (ICLR 2025)

- Learns multiple embeddings per modality to acquire relation-aware modality embeddings
- Integrates the predictions from multiple modalities to achieve joint decisions
- Disentangles the embeddings by minimizing their mutual information

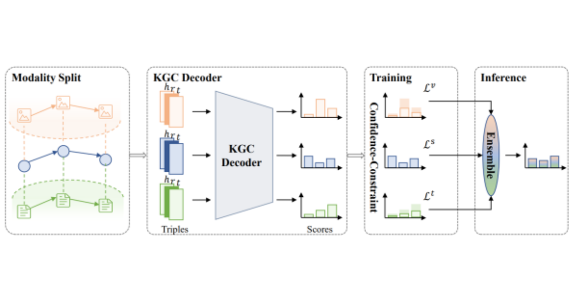


I

MoSE: Modality Split and Ensemble for Multimodal Knowledge Graph Completion

Yu Zhao, Xiangrui Cai*, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang

EMNLP 2022

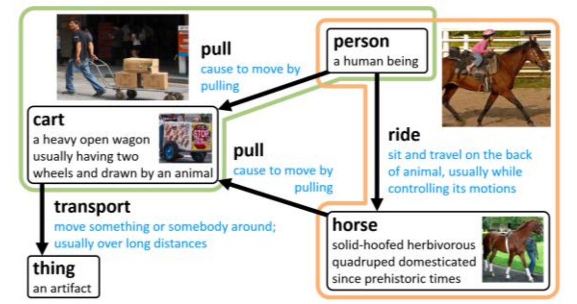


II

VISTA: Visual-Textual Knowledge Graph Representation Learning

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

EMNLP Findings 2023

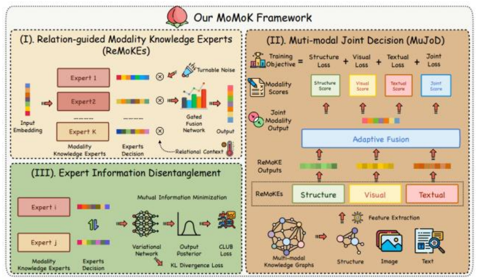


III

Multiple Heads are Better than One: Mixture of Modality Knowledge Experts for Entity Representation Learning

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang*, and Huajun Chen*

ICLR 2025

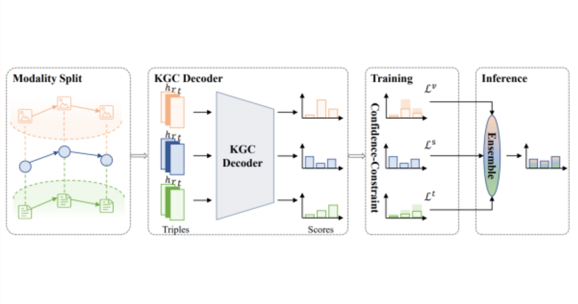


I

MoSE: Modality Split and Ensemble for Multimodal Knowledge Graph Completion

Yu Zhao, Xiangrui Cai*, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang

EMNLP 2022

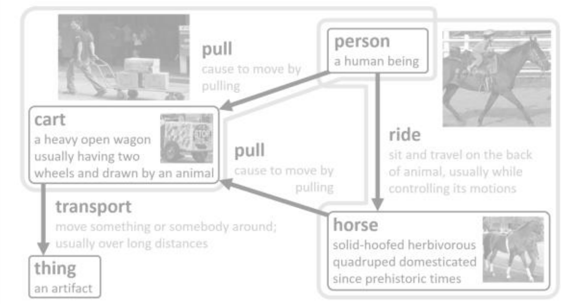


II

VISTA: Visual-Textual Knowledge Graph Representation Learning

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

EMNLP Findings 2023

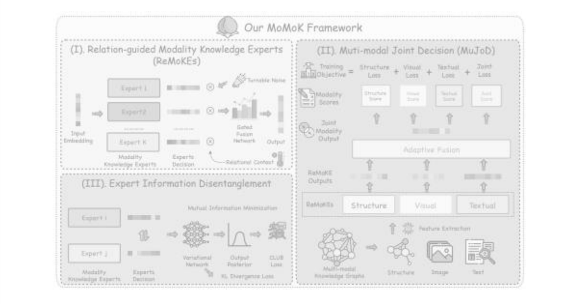


III

Multiple Heads are Better than One: Mixture of Modality Knowledge Experts for Entity Representation Learning

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang*, and Huajun Chen*

ICLR 2025



02 Motivation

- Problem #1: **Modality relation contradiction**
  - Existing methods usually simultaneously represent multiple relations from different modalities only with a single embedding
  - Relation from one modality may contradict that from another modality

- Problem #2: **Modality difference ignorance**
  - Existing methods usually treat the input in different modalities equally and make a unified prediction
  - Different modalities vary in data quality and entity coverage, and should contribute to the final prediction in varying degrees

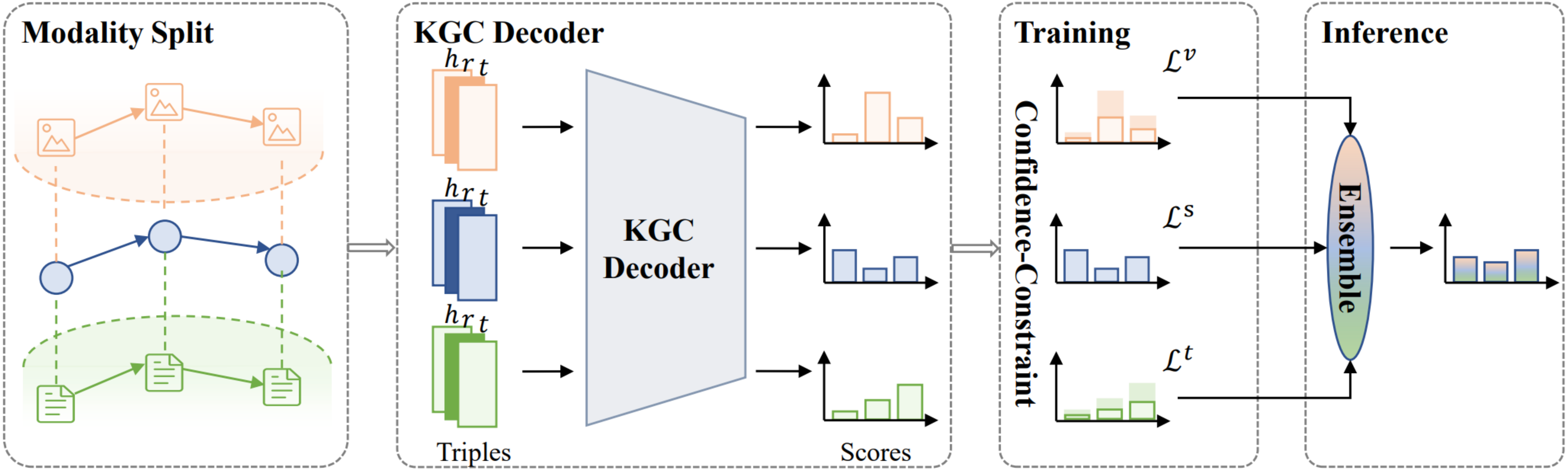
02 Contributions

- Deal with the **modality contradiction of relation representation** and **discuss modality importance** in MKGC task

- Propose a **Mo**dality-**S**plit learning and **E**nsemble inference framework (**MoSE**)
  - Decouples the tight-coupling relation embedding into modality-split ones in the training phase
  - Modulate modality importance adaptively in the inference phase

- MoSE outperforms 9 baseline methods on 3 benchmark multimodal KGC datasets
  - Text modality is a useful component for MKGC rather than visual modality

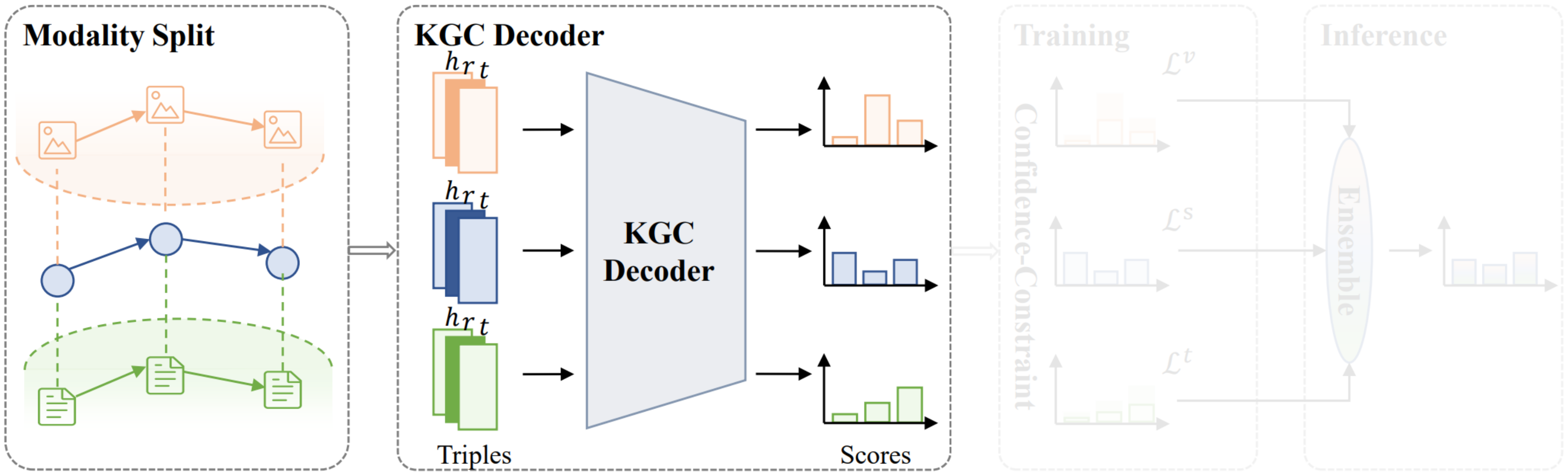
02 Overview of MoSE



02 Modality Split and KGC Decoder

- Utilize **different** relation embeddings for **different** modalities
  - Alleviates modality interference in relation embeddings

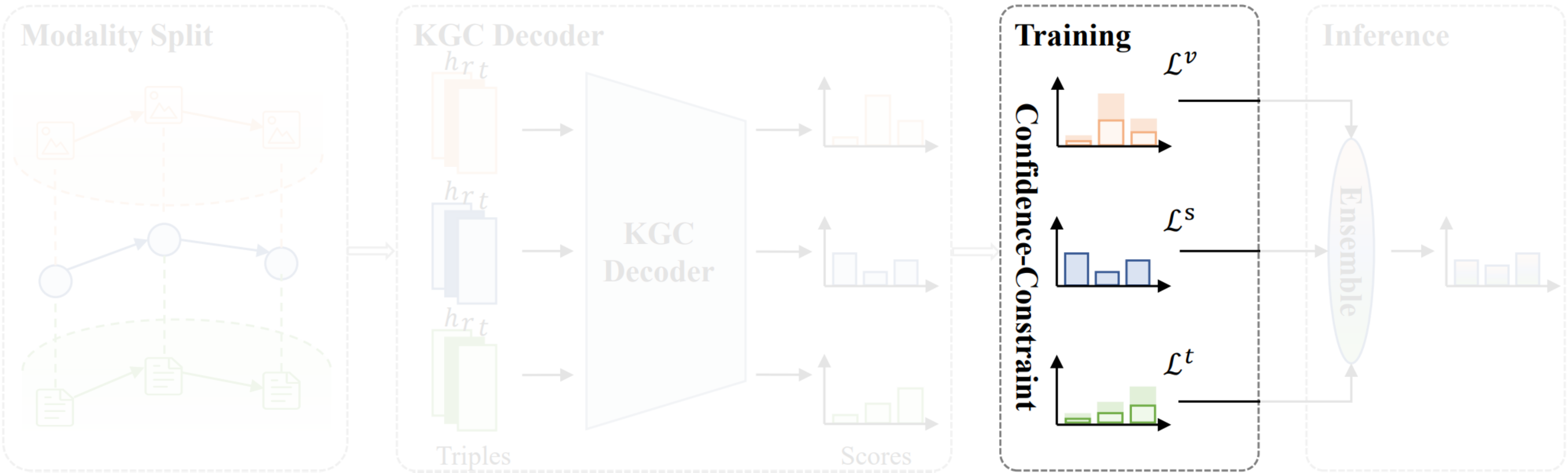
- **Separately compute scores** for each modality
  - Reflects the strengths and limitations of each modality for link prediction



02 Confidence-constraint Training

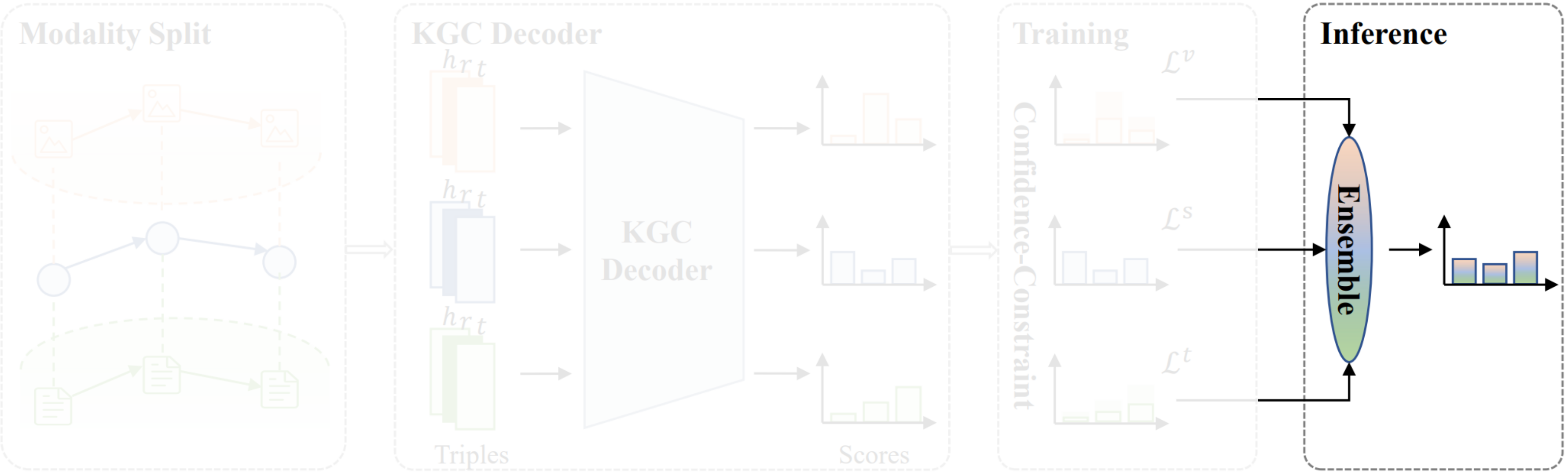
- Visual and textual modalities usually **embody contradictory information** due to **data complexity and diversity,** presenting uncertainty
  - Modality information of entity is not always relevant to the knowledge of a fact triplet

- To ease the uncertainty, the confidence of predictions with visual or textual modality is constrained by adding a temperature parameter



02 Ensemble Inference

- Directly **combines scores from the modalities** to obtain the final score
  - Different weights for different modalities
- MoSE-AI: equal weights for all modalities
- MoSE-BI: relation-specific modality weights
- MoSE-MI: uses an MLP that finds optimal weights based on the scores



02 Experiments

Model	FB15K-237				WN18				WN9			
	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓
Unimodal KGE methods												
TransE	0.198	0.376	0.441	323	0.040	0.745	0.923	357	0.864	0.901	0.917	146
DistMult	0.199	0.301	0.466	512	0.335	0.876	0.940	655	0.531	0.871	0.911	241
ComplEx	0.194	0.297	0.450	546	0.936	0.945	0.947	-	0.901	0.913	0.922	256
RotatE	0.241	0.375	0.533	177	0.942	0.950	0.957	254	0.889	0.906	0.922	175
Multimodal KGE methods												
IKRL (UNION)	0.194	0.284	0.458	298	0.127	0.796	0.928	596	-	-	0.938	21
TransAE	0.199	0.317	0.463	431	0.323	0.835	0.934	352	-	-	0.942	17
RSME	0.242	0.344	0.467	417	0.943	0.951	0.957	223	0.878	0.912	0.923	55
MoSE-AI	0.255	0.376	0.518	135	0.929	0.946	0.962	23	0.840	0.932	0.963	**4**
MoSE-BI	**0.281**	**0.411**	**0.565**	**117**	0.884	0.953	0.972	8	0.831	0.923	0.964	**4**
MoSE-MI	0.268	0.394	0.540	127	**0.948**	**0.962**	**0.974**	**7**	**0.909**	**0.937**	**0.967**	**4**
Pre-trained Language Model methods												
KG-BERT	-	-	0.420	153	0.117	0.689	0.926	58	-	-	-	-
MKGformer	0.256	0.367	0.504	221	0.944	0.961	0.972	28	-	-	-	-

02 Conclusion

- Propose **MoSE**, a novel **modality split learning and ensemble inference framework** for multimodal KGC
  - MoSE decouples modality-shared relation embeddings and performs modality-split representation learning to overcome modality relation contradiction
  - MoSE exploits three ensemble inference techniques to combine the modality-split predictions

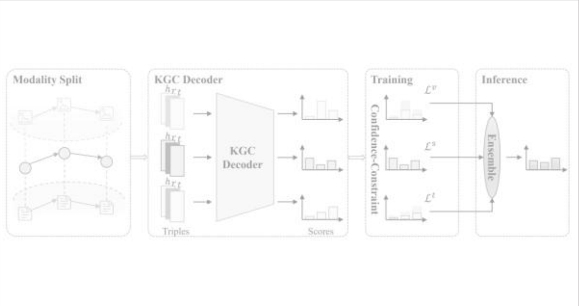
- Experimental results demonstrate that MoSE outperforms state-of-the-art methods for multimodal KGC task on three widely-used datasets

I

MoSE: Modality Split and Ensemble for Multimodal Knowledge Graph Completion

Yu Zhao, Xiangrui Cai*, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang

EMNLP 2022

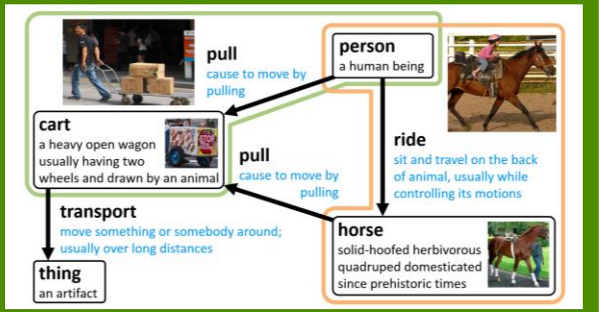


II

VISTA: Visual-Textual Knowledge Graph Representation Learning

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

EMNLP Findings 2023

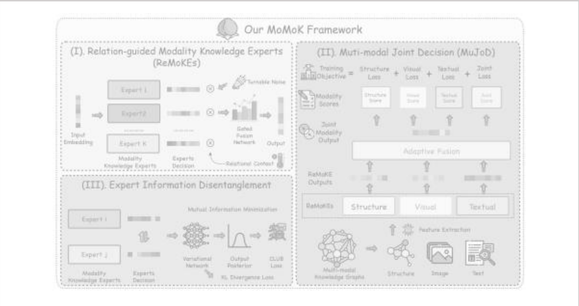


III

Multiple Heads are Better than One: Mixture of Modality Knowledge Experts for Entity Representation Learning

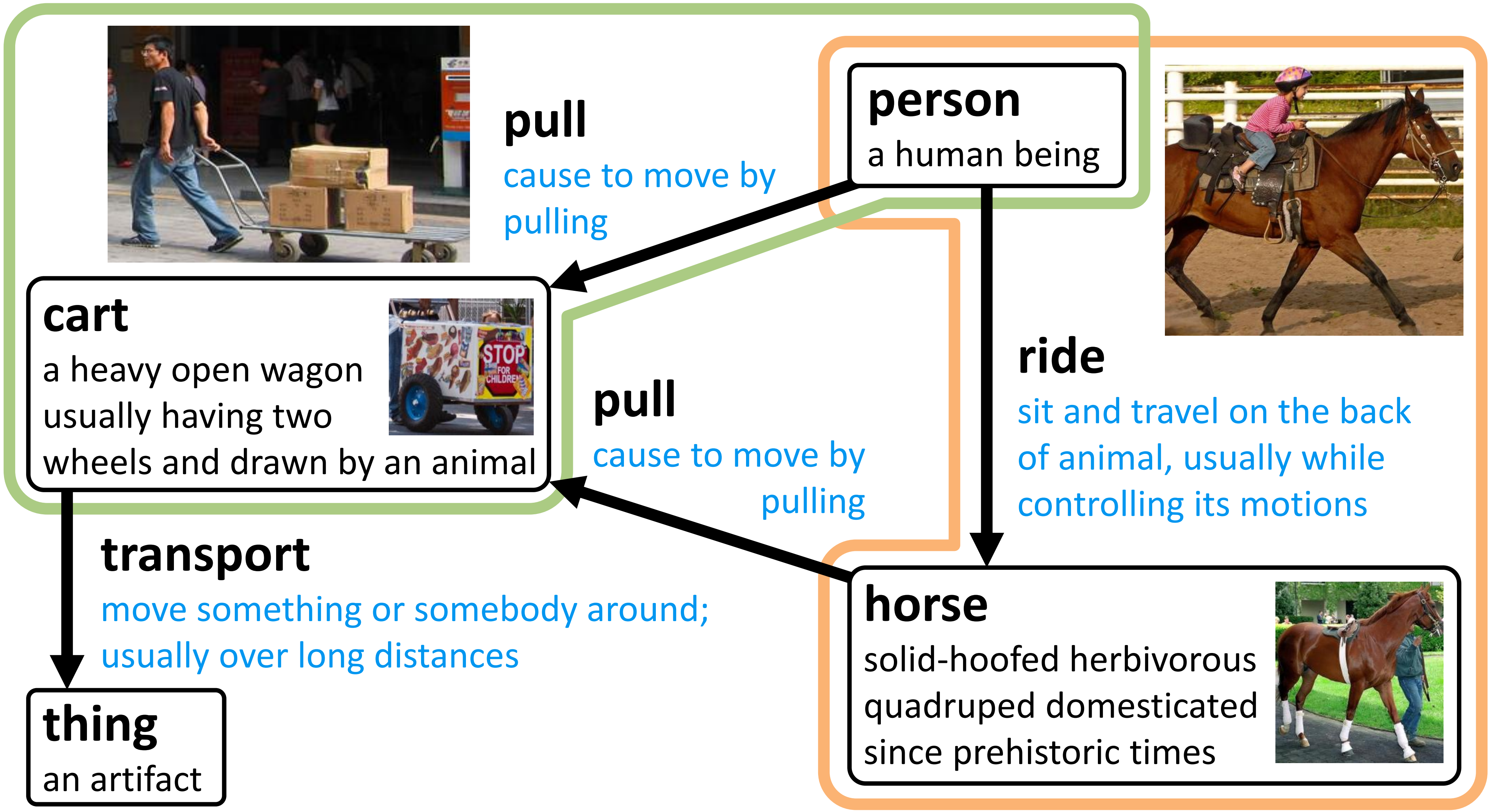
Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang*, and Huajun Chen*

ICLR 2025

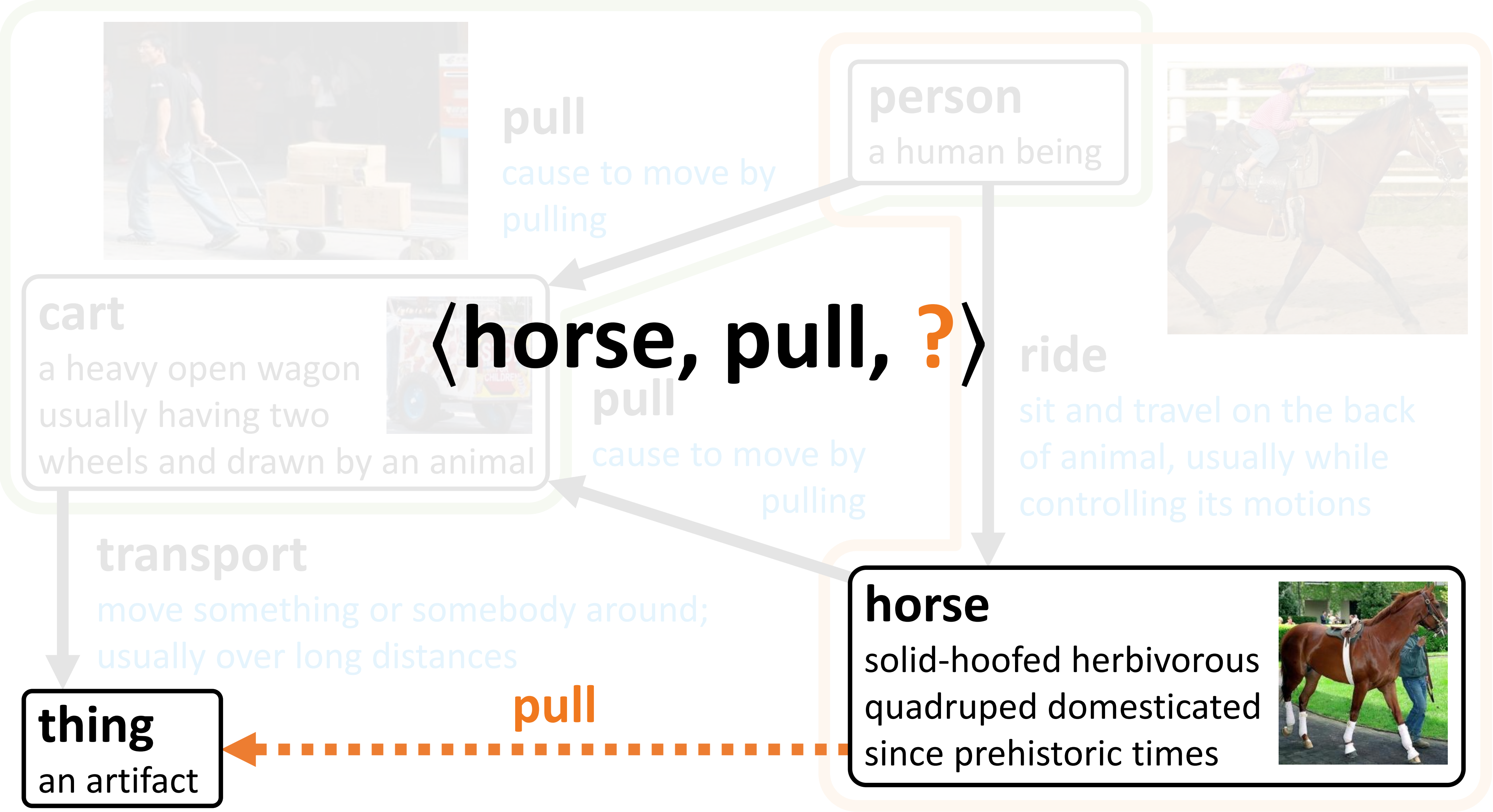


03

Visual-Textual Knowledge Graphs (VTKGs)



03 Link Prediction on VTKGs



03 Contributions

- Define **V**isual-**T**extual **K**nowledge **G**raphs (**VTKGs**)
  - Create two real-world datasets: **VTKG-C** and **VTKG-I**

- **VIS**ual-**T**extu**A**l (**VISTA**) knowledge graph representation learning method
  - VISTA utilizes the **visual and textual features of relations and entities**
  - Define an entity encoder, a relation encoder, and a triplet decoder

- VISTA outperforms **10 different** state-of-the-art knowledge graph completion methods, including multimodal knowledge graph representation learning methods

03 Creating Real-World VTKGs

VRD



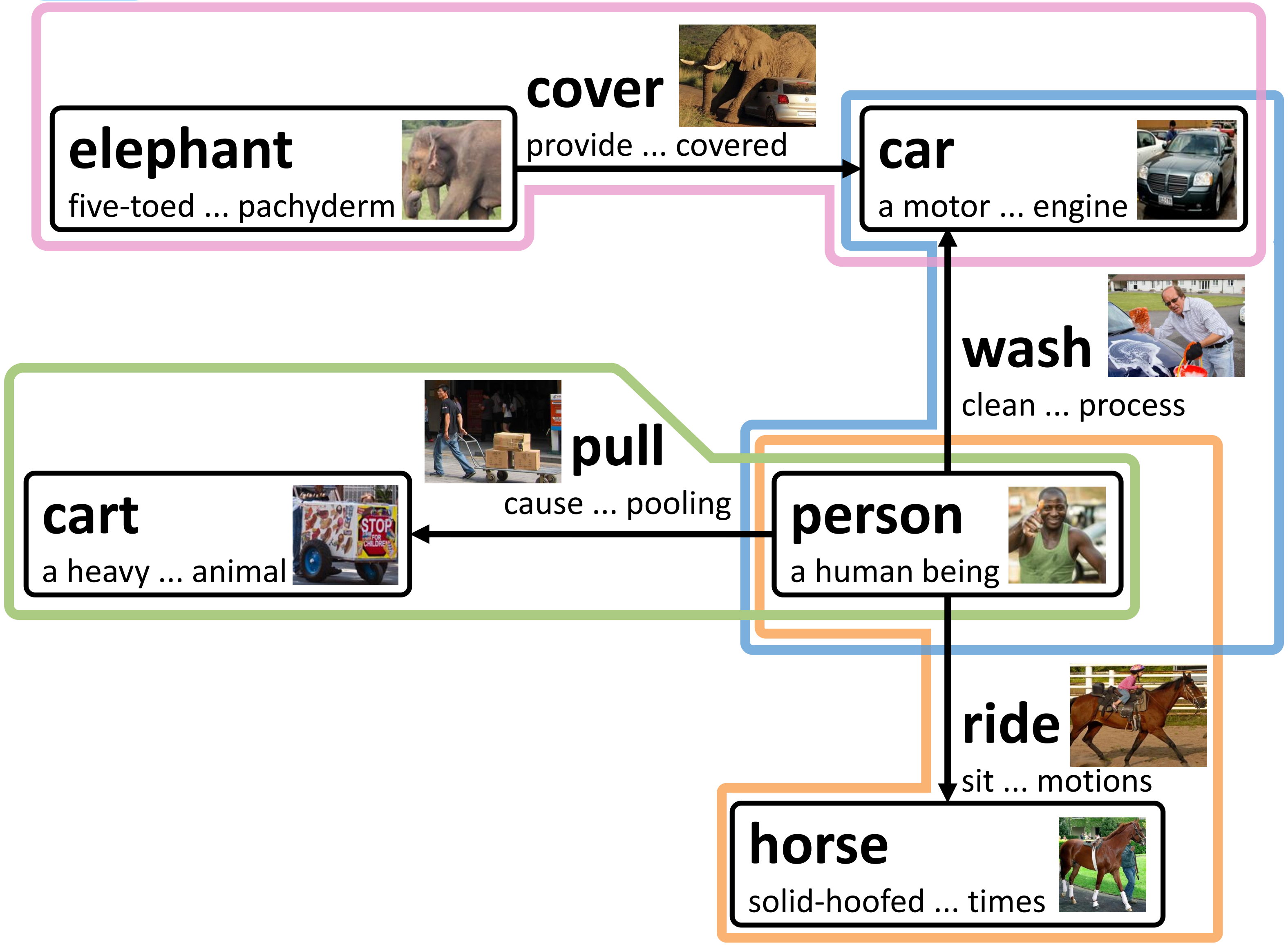
HICO-DET

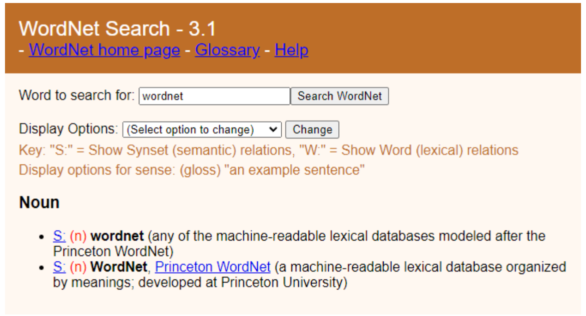


UnRel



03 Creating Real-World VTKGs: VTKG-I





WordNet

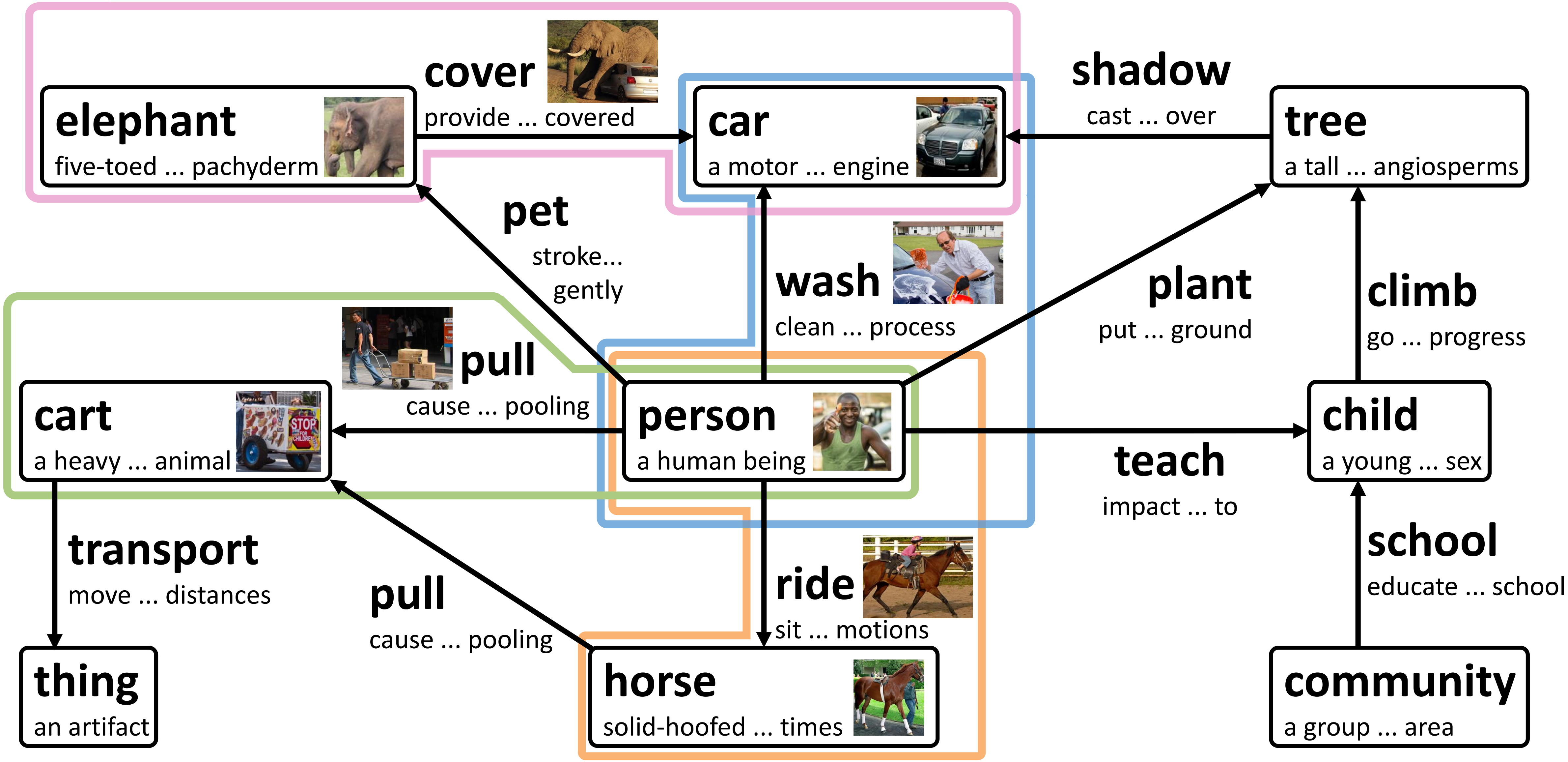
ConceptNet
An open, multilingual knowledge graph

ConceptNet

VisKE

VisKE

03 Creating Real-World VTKGs: VTKG-C



03 Extracting Visual and Textual Features of Entities

Visual Features of horse

ViT-Base → $\mathbf{h}_{\mathrm{vis},1}$

ViT-Base → $\mathbf{h}_{\mathrm{vis},2}$

ViT-Base → $\mathbf{h}_{\mathrm{vis},3}$

ViT-Base → $\mathbf{h}_{\mathrm{vis},4}$

Textual Feature of horse

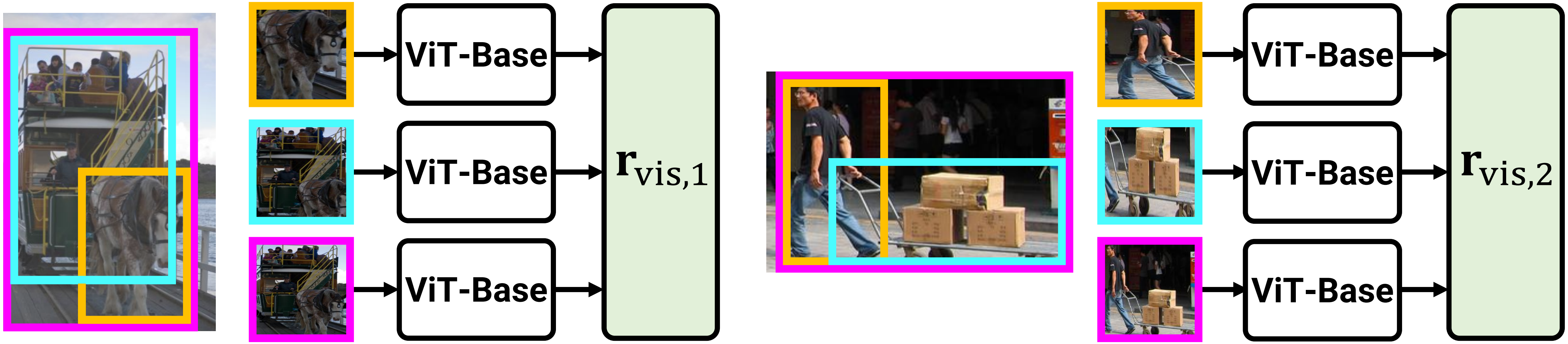
solid-hoofed herbivorous quadruped
domesticated since prehistoric times

→ $\mathrm{BERT}_{\mathrm{BASE}}$ → $\mathbf{h}_{\mathrm{txt}}$

03

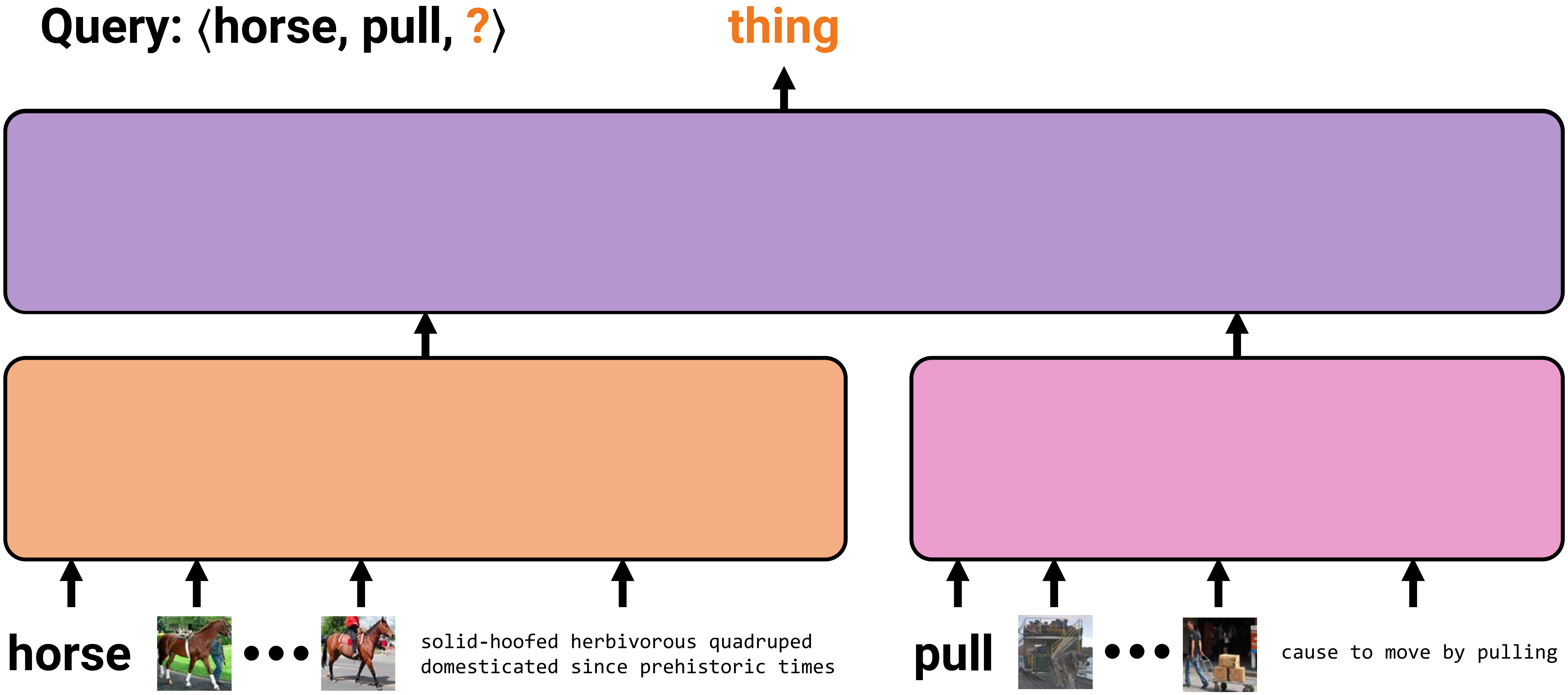
Extracting Visual and Textual Features of Relations

Visual Features of pull

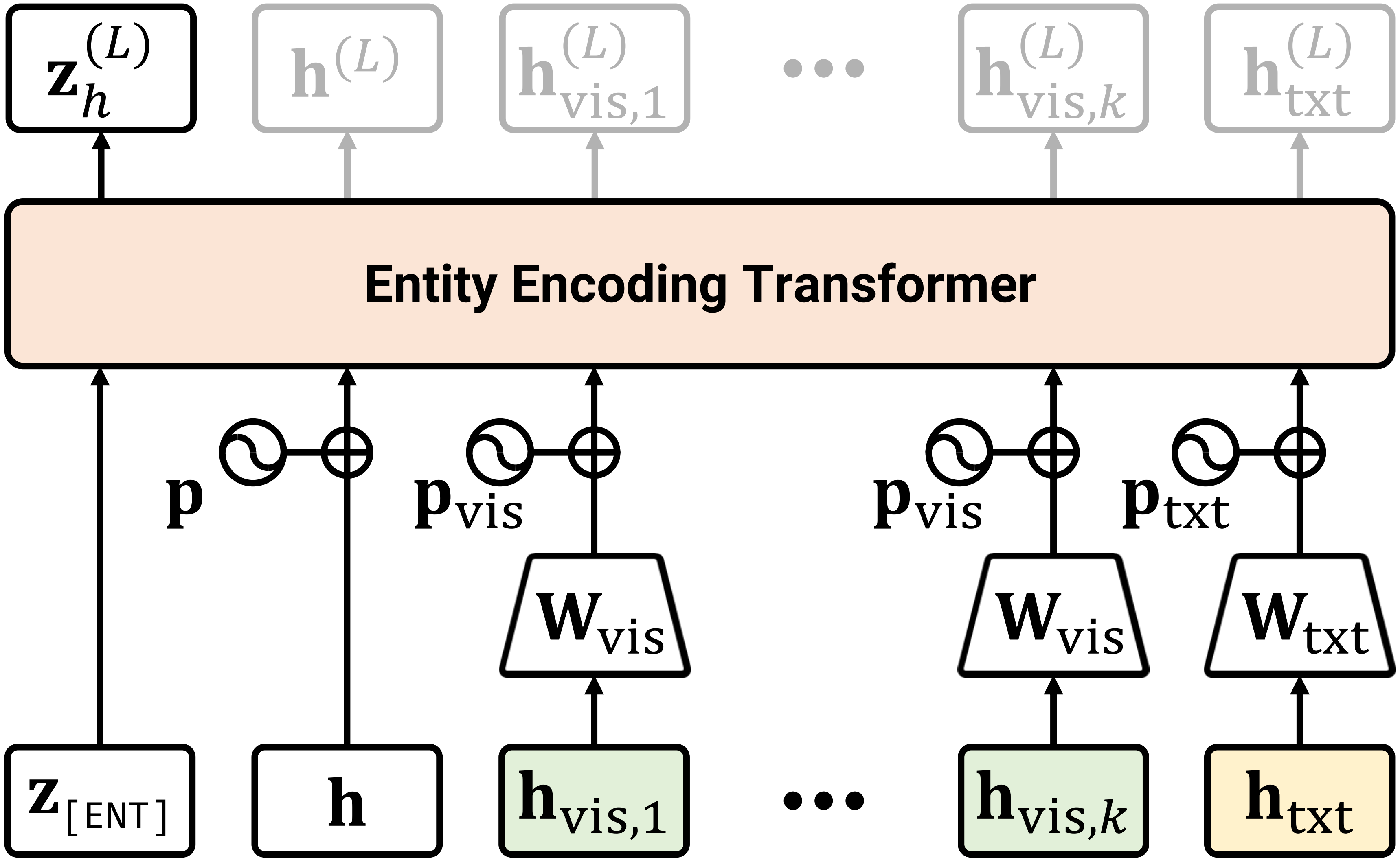


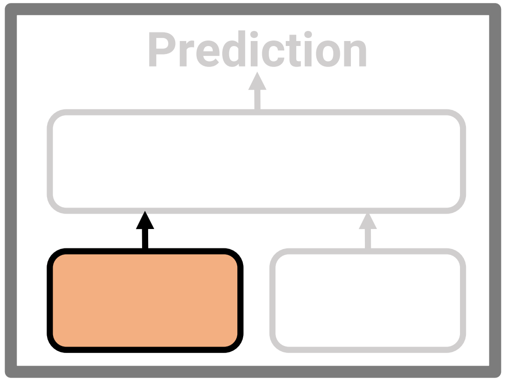
Textual Feature of pull

03 Overview of VISTA

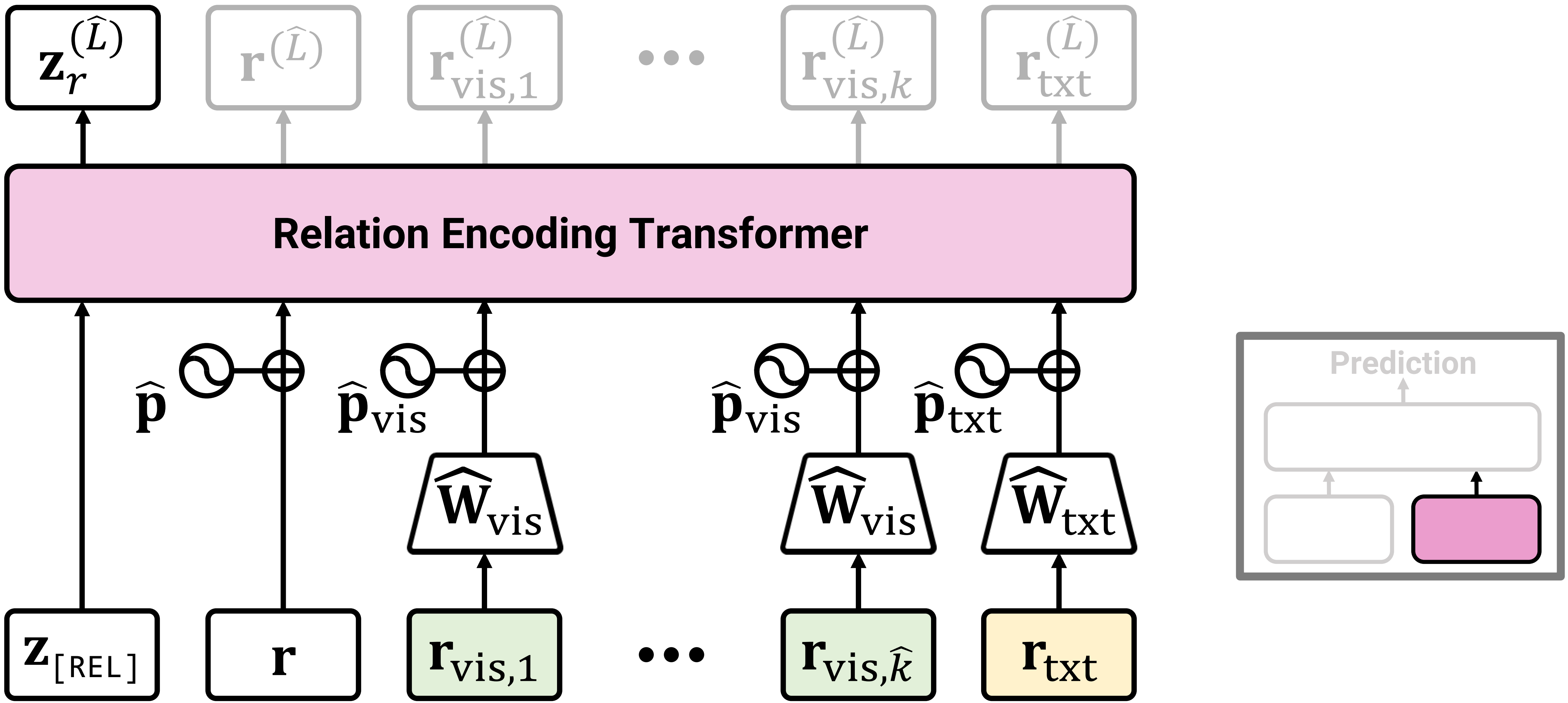


03 Entity Encoder

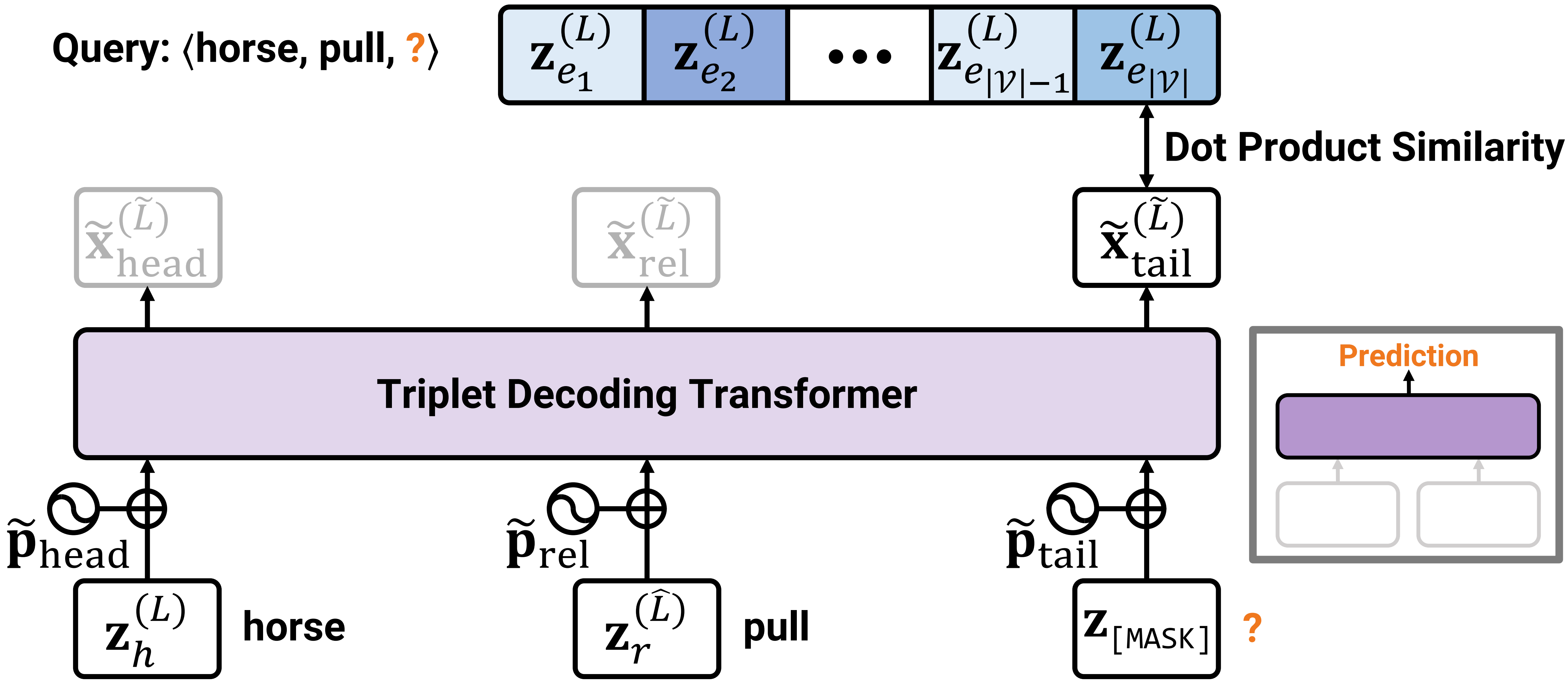




03 Relation Encoder



03 Triplet Decoder



03 Experiments

- Datasets
  - Create two **Visual-Textual Knowledge Graphs (VTKGs)**
    - VTKG-I, VTKG-C
  - Two Benchmark Multimodal Knowledge Graphs
    - WN18RR++ (WN18RR with corrections), FB15K237

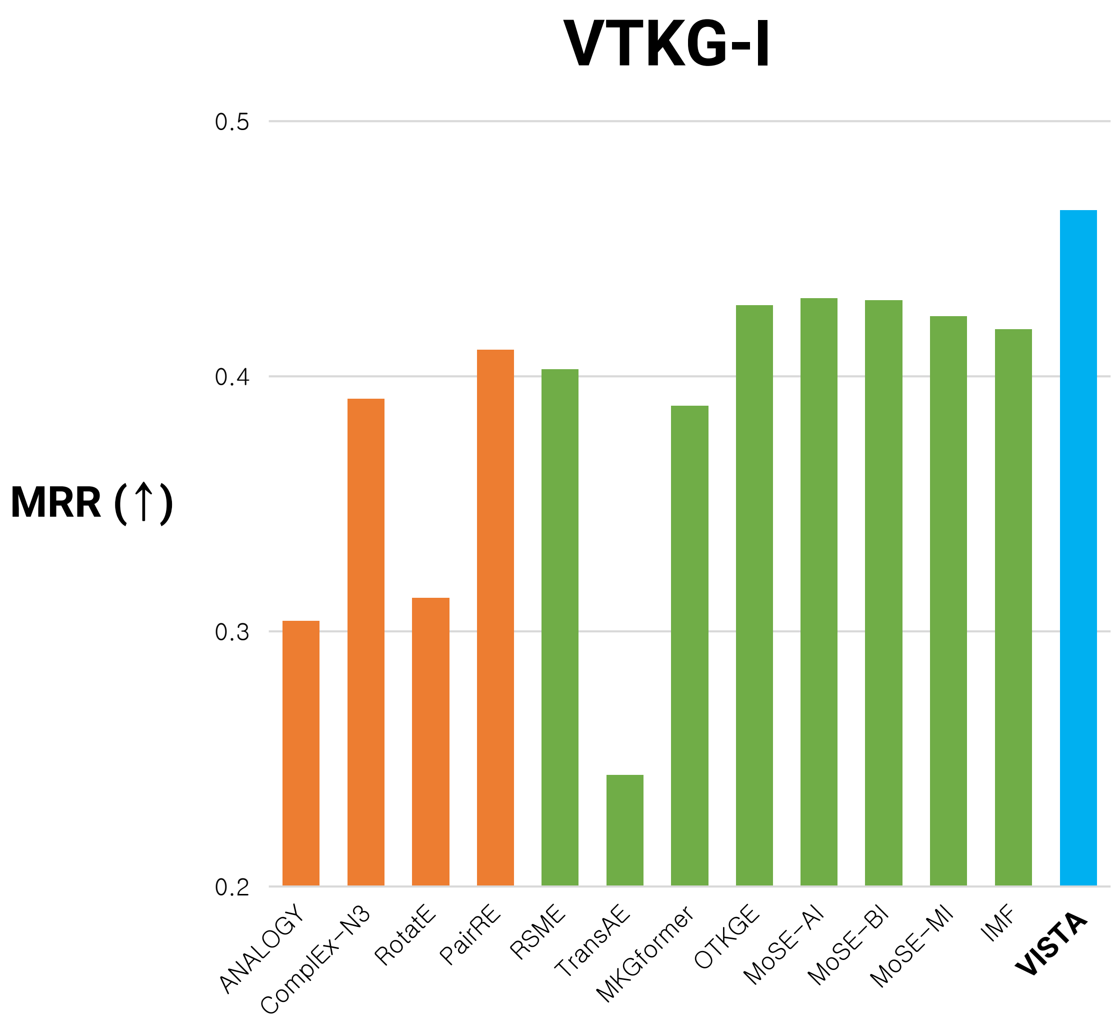
No. of Images No. of Text Descriptions

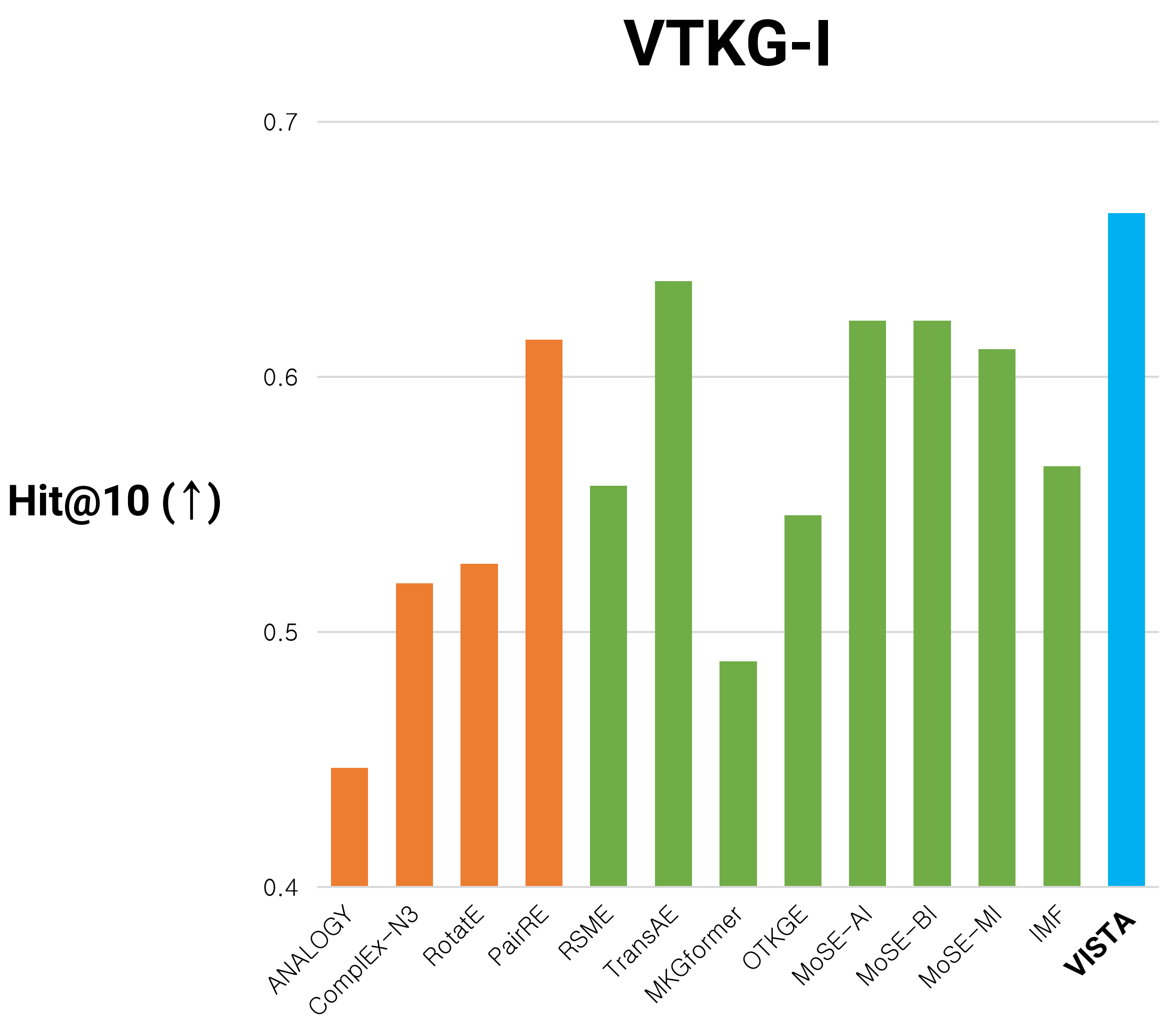
	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{D}\rvert$
VTKG-I	181	217	1,316	390,658	383
VTKG-C	43,267	2,731	111,491	461,007	45,401
WN18RR++	41,105	11	93,003	70,349	41,105
FB15K237	14,541	237	310,116	145,944	14,515

03 Experiments

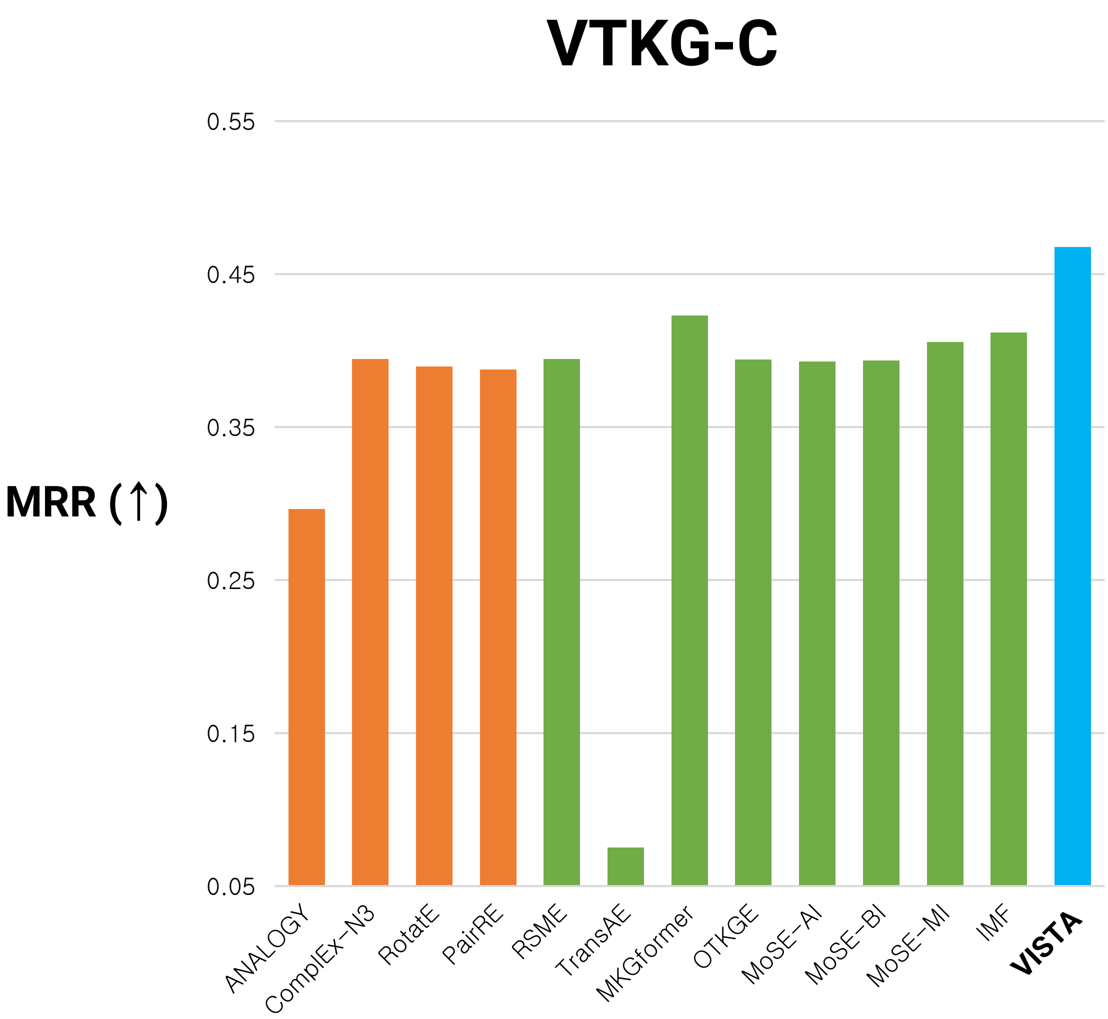
- Comparison with **10 baseline methods**
  - Knowledge Graph Embedding Methods
    - ANALOGY (ICML 2017)
    - ComplEx-N3 (ICML 2018)
    - RotatE (ICLR 2019)
    - PairRE (ACL 2021)
  - Multimodal Knowledge Graph Representation Learning Methods
    - RSME (MM 2021)
    - TransAE (IJCNN 2019)
    - MKGformer (SIGIR 2022)
    - OTKGE (NeurIPS 2022)
    - MoSE (EMNLP 2022)
    - IMF (TheWebConf 2023)

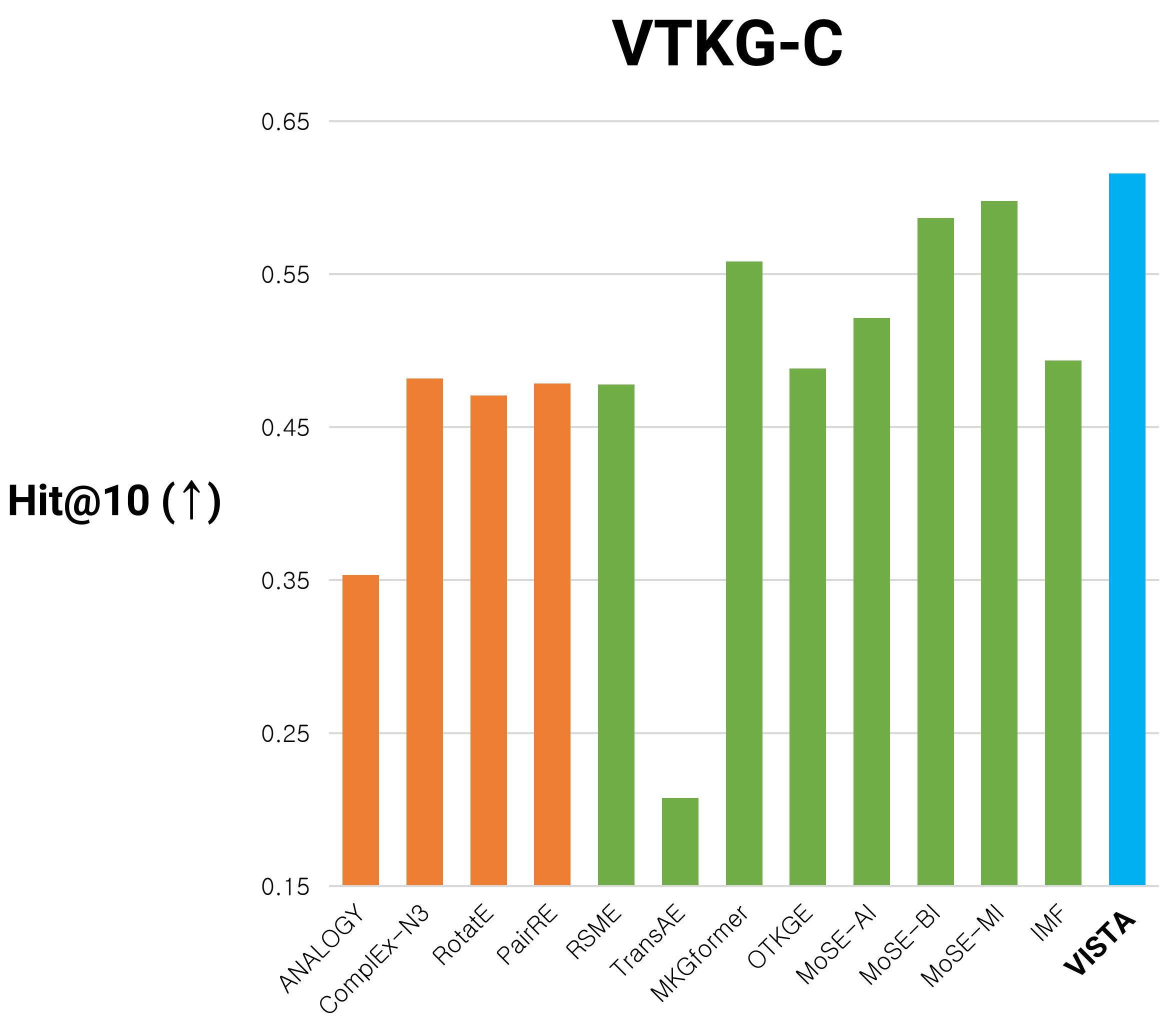
03 Link Prediction Performance



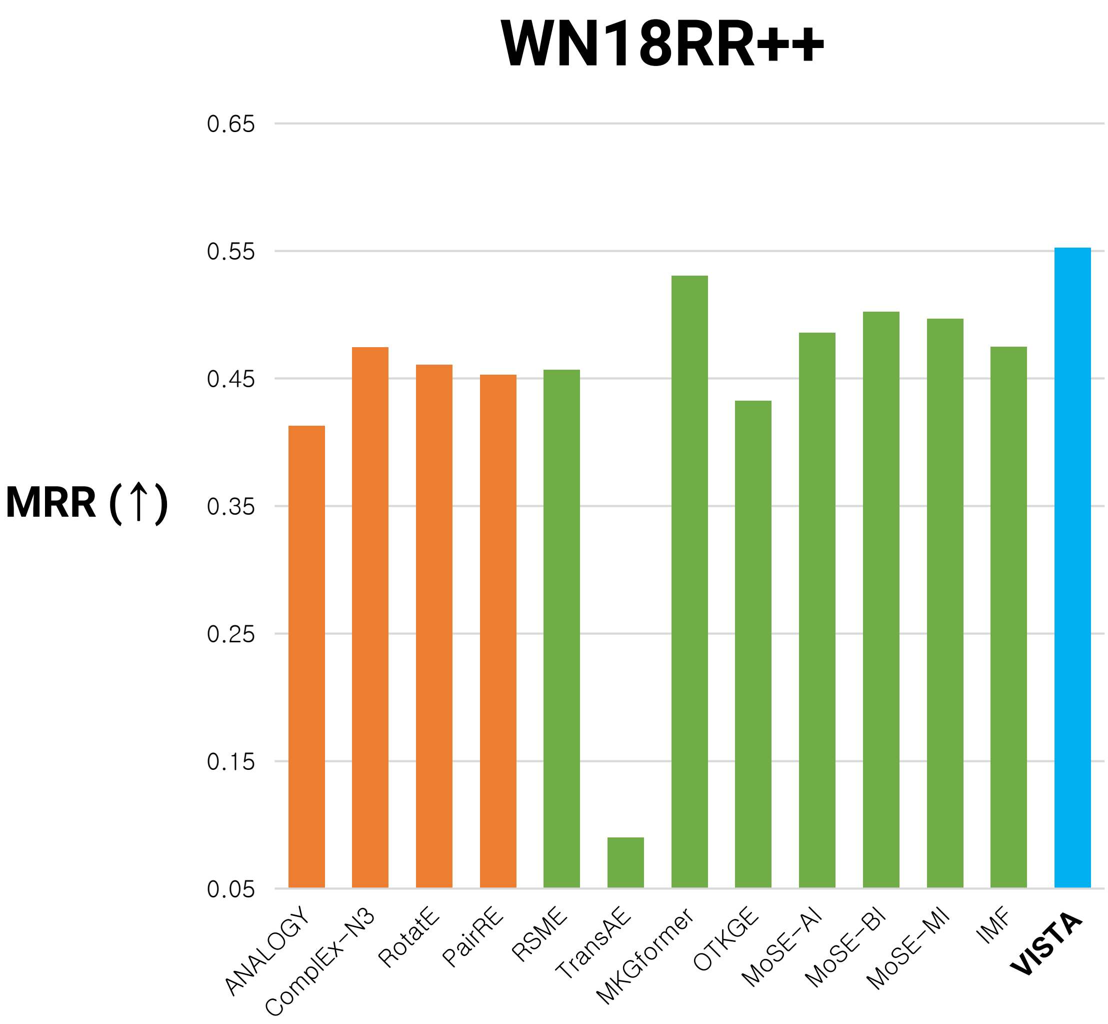


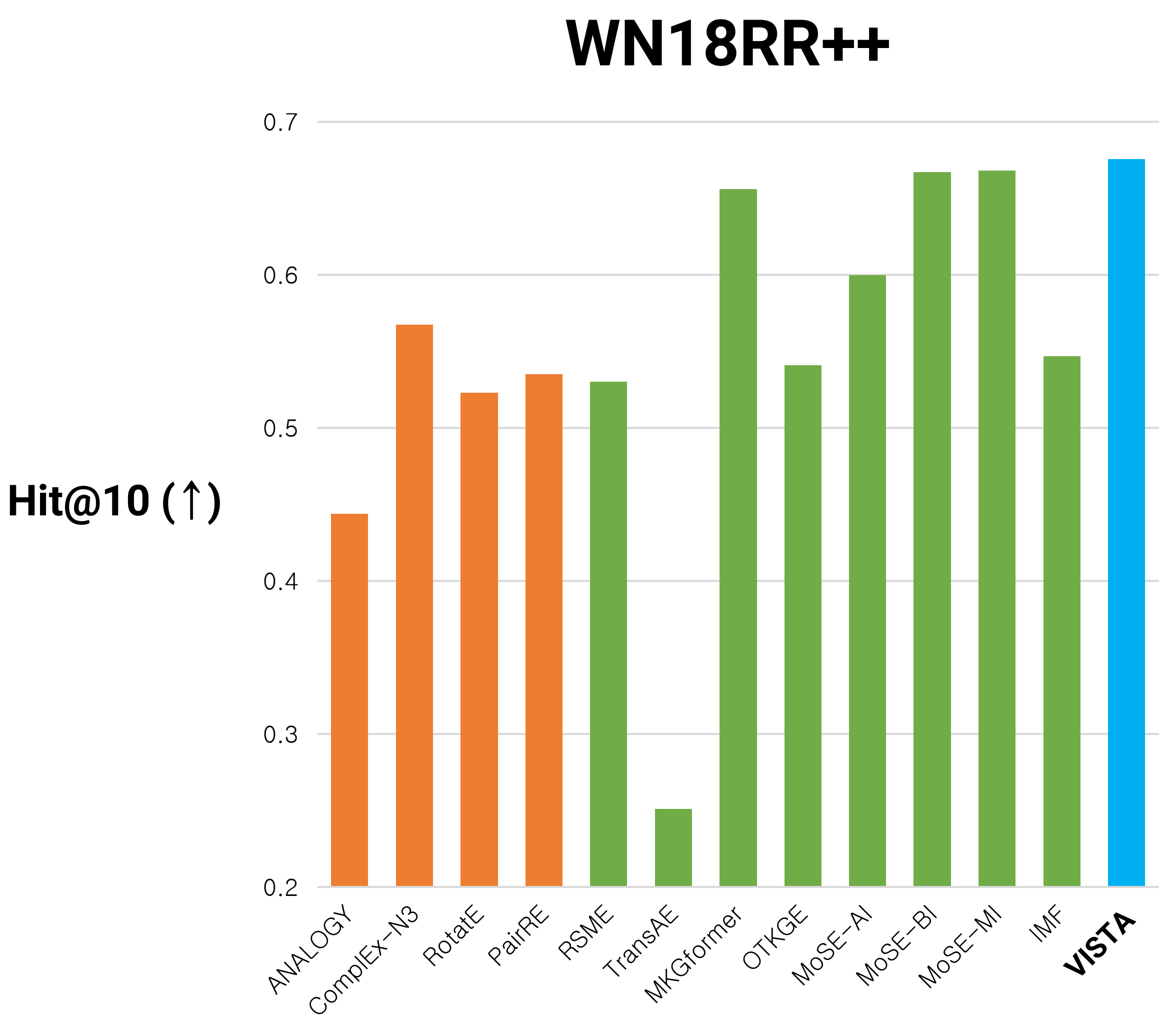
03 Link Prediction Performance



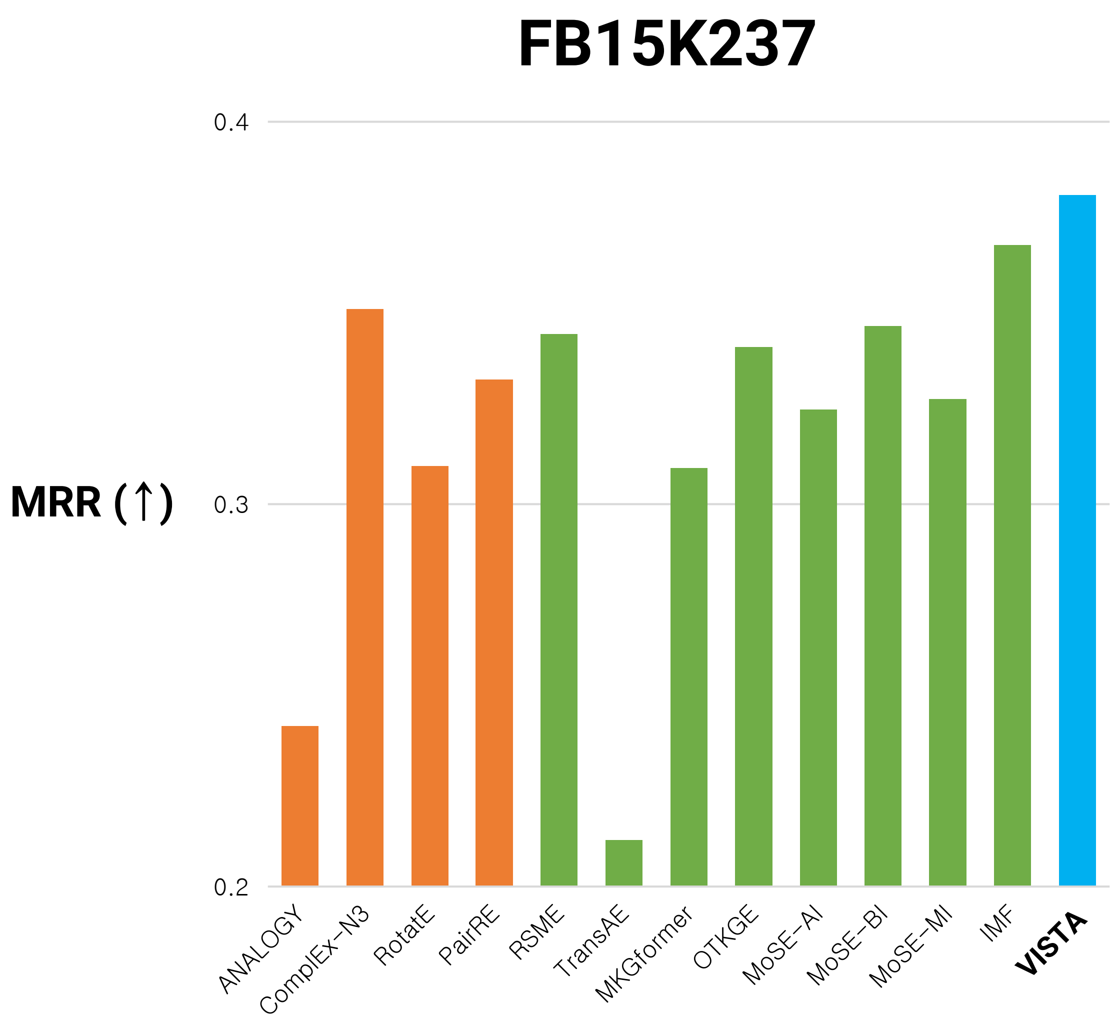


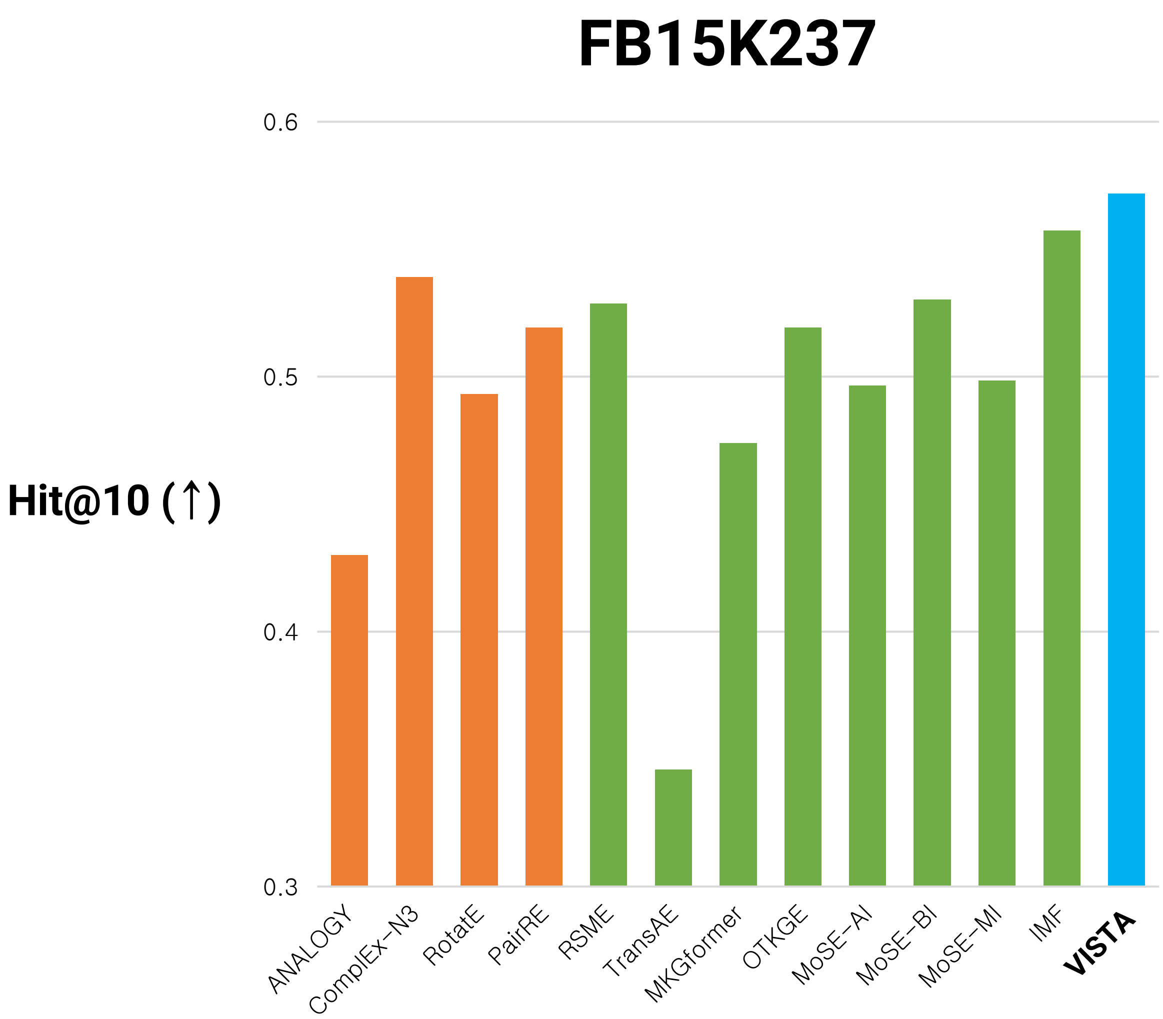
03 Link Prediction Performance





03 Link Prediction Performance





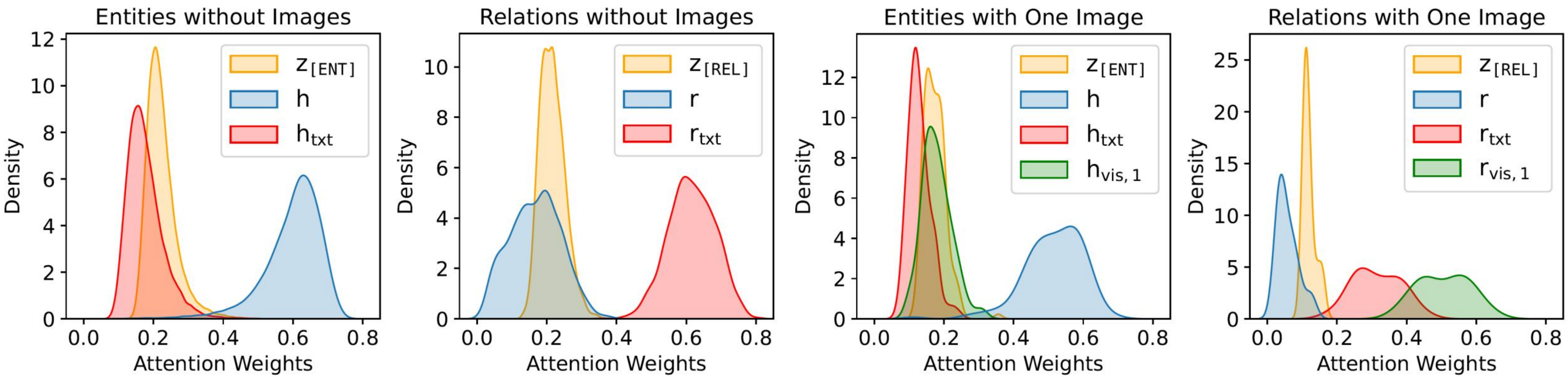
03 Top Similar Entities & Relations

- BERT returns **abstract concepts**; ViT returns **visually expressible concepts**.
- VISTA returns the most semantically close entities and relations to the queries by utilizing **both texts and images**.

Query		BERT	ViT	VISTA
dark_red	1	incense	leisure_wear	orange
	2	coloring	sportswear	red
	3	buffer	sweatshirt	crimson
have	1	move	straddle	keep
	2	influence	hop_on	hold
	3	begin	inspect	incorporate

03 Attention Weights

- When images are not given, **learnable vectors** have relatively high attention weights in entities whereas **textual features** play the crucial role in relations.
- When an image is given, **learnable vectors** still has high importance in entities whereas **visual features** tend to have high contributions in relations.



03 Conclusion

- **V**isual-**T**extual **K**nowledge **G**raphs (**VTKGs**)
  - Visually expressible triplets are augmented by images
  - Both entities and relations have textual descriptions

- Propose **VIS**ual-**T**extu**A**l (**VISTA**) knowledge graph representation learning method to solve knowledge graph completion problems in real-world VTKG datasets

- VISTA takes into account the visual and textual features of entities and relations

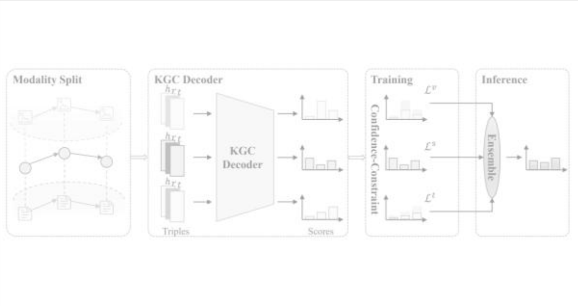
- VISTA substantially outperforms 10 different state-of-the-art methods

I

MoSE: Modality Split and Ensemble for Multimodal Knowledge Graph Completion

Yu Zhao, Xiangrui Cai*, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang

EMNLP 2022

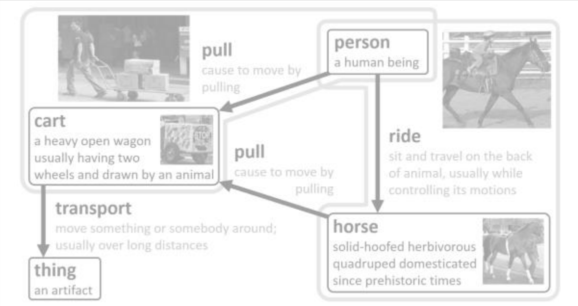


II

VISTA: Visual-Textual Knowledge Graph Representation Learning

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

EMNLP Findings 2023

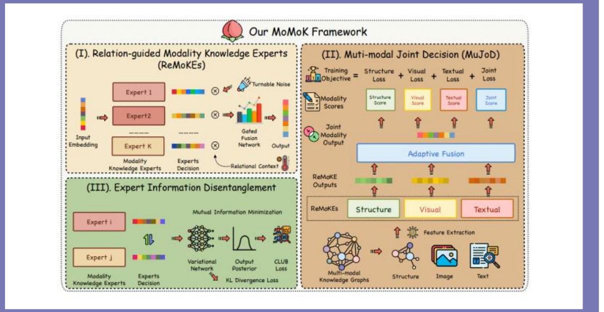


III

Multiple Heads are Better than One: Mixture of Modality Knowledge Experts for Entity Representation Learning

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang*, and Huajun Chen*

ICLR 2025



04 Motivation

- Existing multimodal KGC methods typically employ a fusion module to integrate the information from different modalities to obtain **joint entity embeddings**
  - The entity embeddings are then mapped into a scalar score along with the relation embeddings as a basis for assessing the plausibility of a triplet

- These approaches **overlook the information diversity** in both **inter-modality** and **intra-modality**
  - Different modalities can represent various aspects of entity information
  - Information within the same modality can also play different roles depending on the relational context

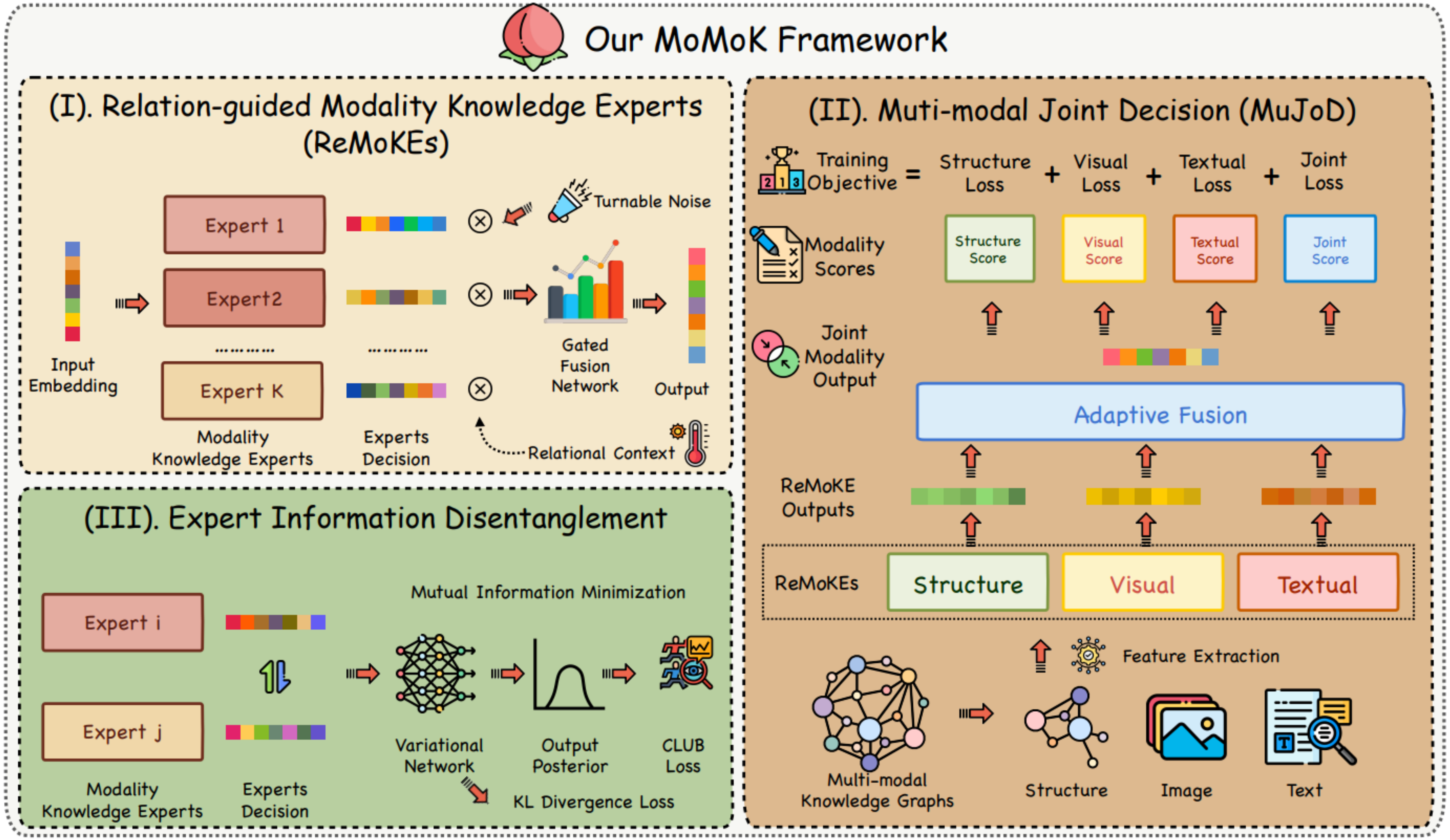
04 Contributions

- Address the problems in modality information utilization by multimodal KGC models and propose **MoMoK**, a **M**ixture **of** **Mo**dality **K**nowledge experts framework
  - MoMoK consists of relational-guided modality experts and multi-modal joint decision

- Examine the **learning of different modality experts** through the lens of **mutual information estimation**, and propose to **decouple the expert information** within each modality using **mutual information comparison estimation**

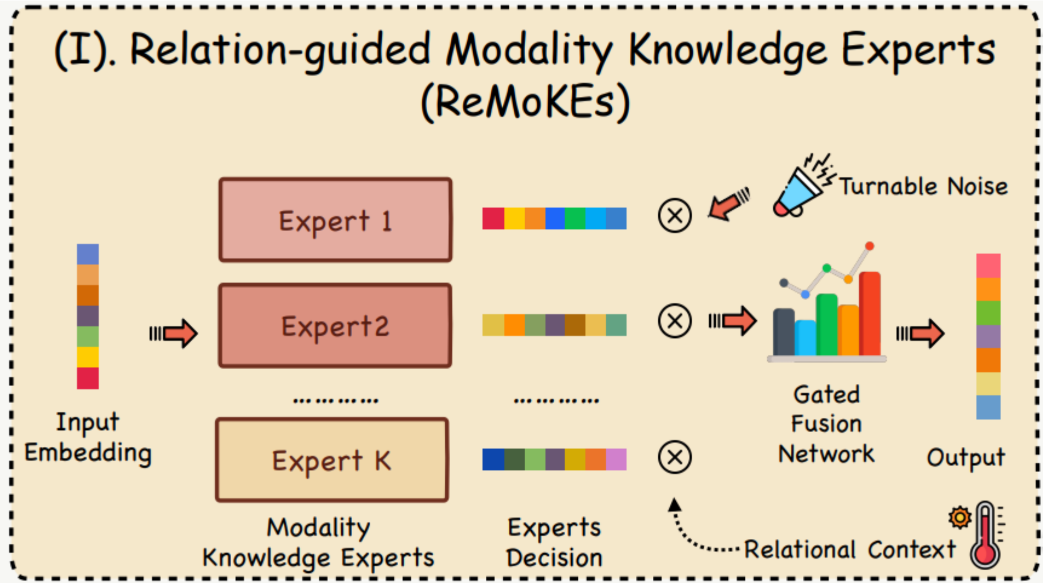
- Conduct experiments against 20 baselines on 4 multimodal KGC benchmarks
  - MoMoK achieves state-of-the-art performance

04 Overview of MoMoK



04 Relation-guided Modality Knowledge Experts

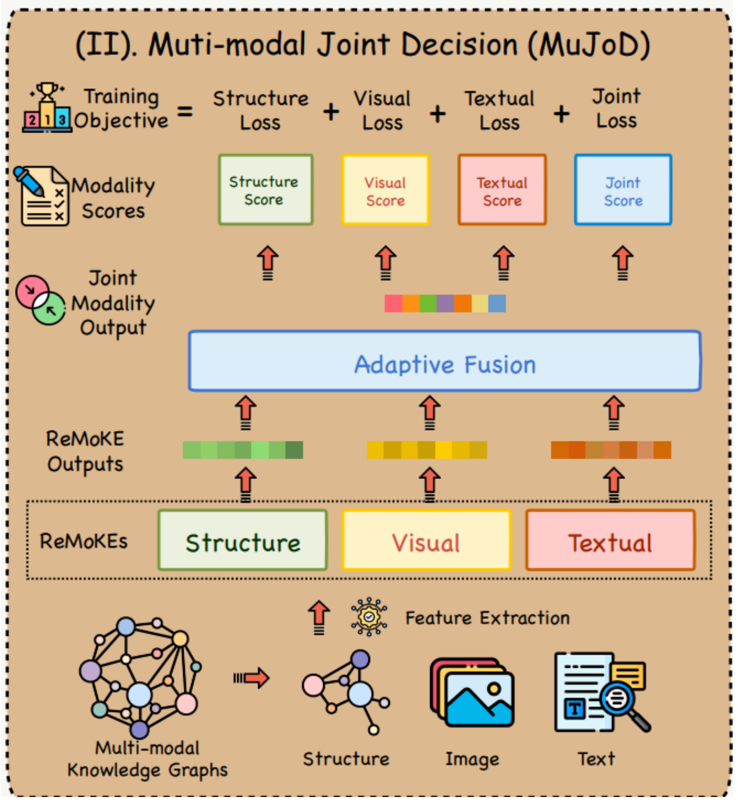
- Designed to learn the embedding of **different perspectives in intra-modalities**
  - Learn multi-perspective embeddings of each entity for each modality
- Compute **relation-specific entity embedding** for each modality by combining the embeddings with relation-specific weights



04 Multi-modal Joint Decision

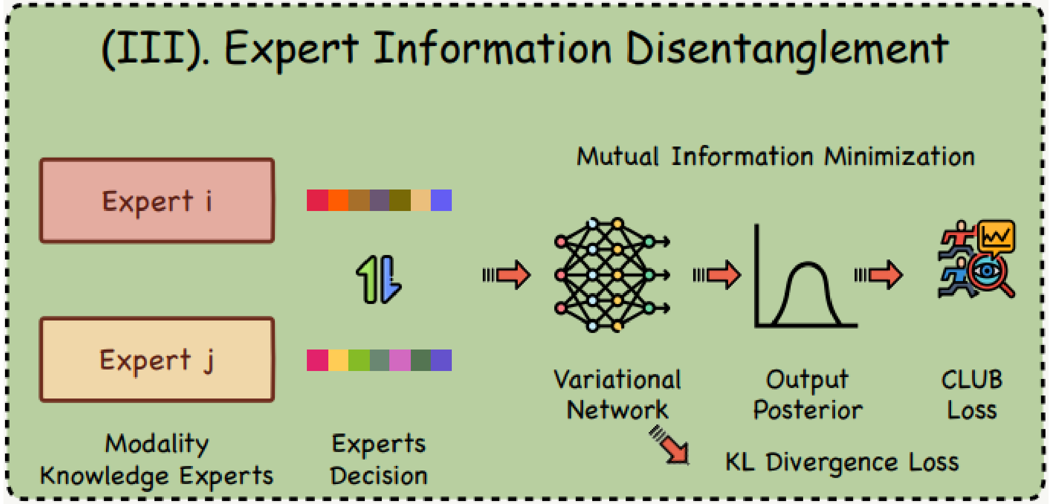
- Perform multimodal entity embedding fusion by **learning a group of adaptive weights** for each entity
  - Aggregates information from all modalities
  - The resulting embedding is considered as **"joint" modality**

- Compute triplet plausibility from **each modality's perspective** and add them to compute the final score
  - During training, the loss of each modality is computed separately, and then directly combined to derive the final loss



04 Expert Information Disentanglement

- Designed to allow the model to **learn multi-perspective embeddings** guided by the relational context
- Disentangle the experts' decisions in each modality by **minimizing the mutual information** between the multi-perspective embeddings for each modality
  - The disentangle loss is combined with prediction losses to yield the final loss



04 Experiments

Model	MKG-W		MKG-Y		DB15K				KVC16K			
	MRR	Hit@1	MRR	Hit@1	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
Uni-modal KGC Methods												
TransE	29.19	21.06	30.73	23.45	24.86	12.78	31.48	47.07	8.54	0.64	10.97	23.42
DistMult	20.99	15.93	25.04	19.33	23.03	14.78	26.28	39.59	6.37	3.03	6.11	12.61
ComplEx	24.93	19.09	28.71	22.26	27.48	18.37	31.57	45.37	12.85	7.48	13.79	23.18
RotatE	33.67	26.80	34.95	29.10	29.28	17.87	36.12	49.66	14.33	8.25	15.37	26.17
PairRE	34.40	28.24	32.01	25.53	31.13	21.62	35.91	49.30	-	-	-	-
TuckER	29.59	23.93	37.05	34.59	33.86	25.34	37.91	50.38	15.90	9.79	17.24	27.58
Multi-modal KGC Methods												
IKRL	32.36	26.11	33.22	30.37	26.82	14.09	34.93	49.09	11.11	5.42	11.46	22.39
TBKGC	31.48	25.31	33.99	30.47	28.40	15.61	37.03	49.86	5.39	0.35	5.04	15.52
TransAE	30.00	21.23	28.10	25.31	28.09	21.25	31.17	41.17	10.81	5.31	11.34	21.89
MMKRL	30.10	22.16	36.81	31.66	26.81	13.85	35.07	49.39	8.78	3.89	8.99	18.34
RSME	29.23	23.36	34.44	31.78	29.76	24.15	32.12	40.29	12.31	7.14	13.21	22.05
VBKGC	30.61	24.91	37.04	33.76	30.61	19.75	37.18	49.44	14.66	8.28	15.81	27.04
OTKGE	34.36	28.85	35.51	31.97	23.86	18.45	25.89	34.23	8.77	5.01	9.31	15.55
MoSE*	33.34	27.78	36.28	33.64	28.38	21.56	30.91	41.67	8.81	4.75	9.46	16.40
IMF*	34.50	28.77	35.79	32.95	32.25	24.20	36.00	48.19	12.01	7.42	12.82	21.01
QEB	32.38	25.47	34.37	29.49	28.18	14.82	36.67	51.55	12.06	5.57	13.03	25.01
VISTA	32.91	26.12	30.45	24.87	30.42	22.49	33.56	45.94	11.89	6.97	12.66	21.27
AdaMF	34.27	27.21	**38.06**	33.49	32.51	21.31	39.67	51.68	15.26	8.56	16.71	28.29
Negative Sampling Methods												
MANS	30.88	24.89	29.03	25.25	28.82	16.87	36.58	49.26	10.42	5.21	11.01	20.45
MMRNS	35.03	28.59	35.93	30.53	32.68	23.01	37.86	51.01	13.31	7.51	14.19	24.68
MoMoK	**35.89**	**30.38**	37.91	**35.09**	**39.57**	**32.38**	**43.45**	**54.14**	**16.87**	**10.53**	**18.26**	**29.20**
Improvements	+2.5%	+4.2%	-	+3.9%	+21.1%	+33.8%	+9.5%	+4.8%	+10.6%	+23.0%	+9.3%	+3.21%

04 Conclusion

- Propose a multimodal KGC framework called **MoMoK** to learn modality features in **diverse perspectives** from the raw modality information of entities
  - MoMoK learns modality features with relation guidance and integrates the multi-modal information through modality knowledge experts

- Expert networks are **decoupled** and enhance the model's expressive capability through the comparative estimation of mutual information

- Experimental results show MoMoK achieve new state-of-the-art results

05 References

- Some slides are made based on the following references.

  - R. Xie et al., "Image-embodied Knowledge Representation Learning", IJCAI, 2017.
  - A. Kristiadi et al., "Incorporating Literals into Knowledge Graph Embeddings", ISWC, 2019.
  - M. Wang et al., "Is Visual Context Really Helpful for Knowledge Graph? A Representation Learning Perspective", MM, 2021.
  - X. Chen et al., "Hybrid Transformer with Multi-level Fusion for Multimodal Knowledge Graph Completion", SIGIR, 2022.
  - Z. Cao et al., "OTKGE: Multi-modal Knowledge Graph Embeddings via Optimal Transport", NeurIPS, 2022.
  - Y. Zhao et al., "MoSE: Modality Split and Ensemble for Multimodal Knowledge Graph Completion", EMNLP, 2022.
  - X. Li et al., "IMF: Interactive Multimodal Fusion Model for Link Prediction", TheWebConf, 2023.
  - J. Lee et al., "VISTA: Visual-Textual Knowledge Graph Representation Learning", EMNLP Findings, 2023.
  - Y. Zhang et al., "Multiple Heads are Better than One: Mixture of Modality Knowledge Experts for Entity Representation Learning", ICLR, 2025.