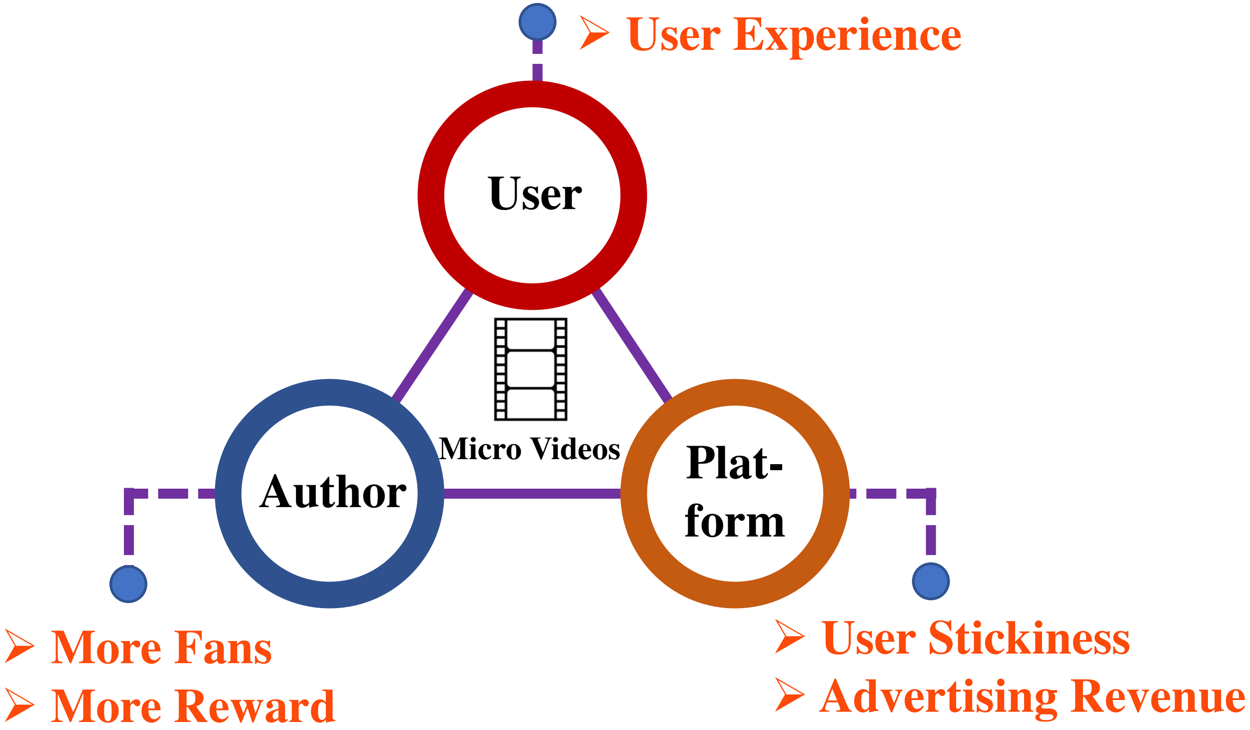
# A Unified Model for Video Understanding and Knowledge Embedding with Heterogeneous Knowledge Graph Dataset

Jiaxin Deng[1], Dong Shen[2], Haojie Pan[2], Xiangyu Wu[2], Ximan Liu[2],
Gaofeng Meng[1*], Fan Yang[2], Tingting Gao[2], Ruiji Fu[2], Zhongyuan Wang[2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]MMU KuaiShou Inc.
[*]Corresponding Author

# Background: Micro-Video Understanding

**Rapid Development of Various Information-sharing Platforms on the Internet**



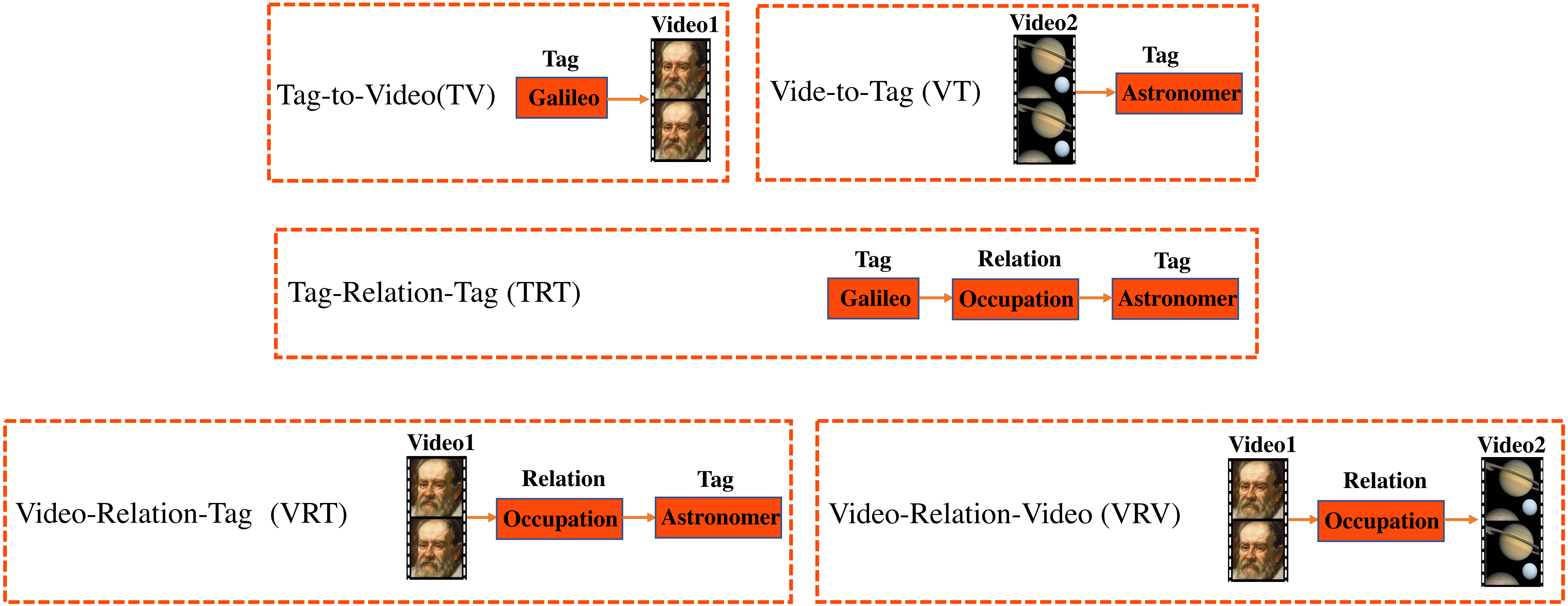


**The Screenshot of the Kuaishou APP**

# Motivation



## ➢ Combining Video Understanding with Knowledge Graph Embedding

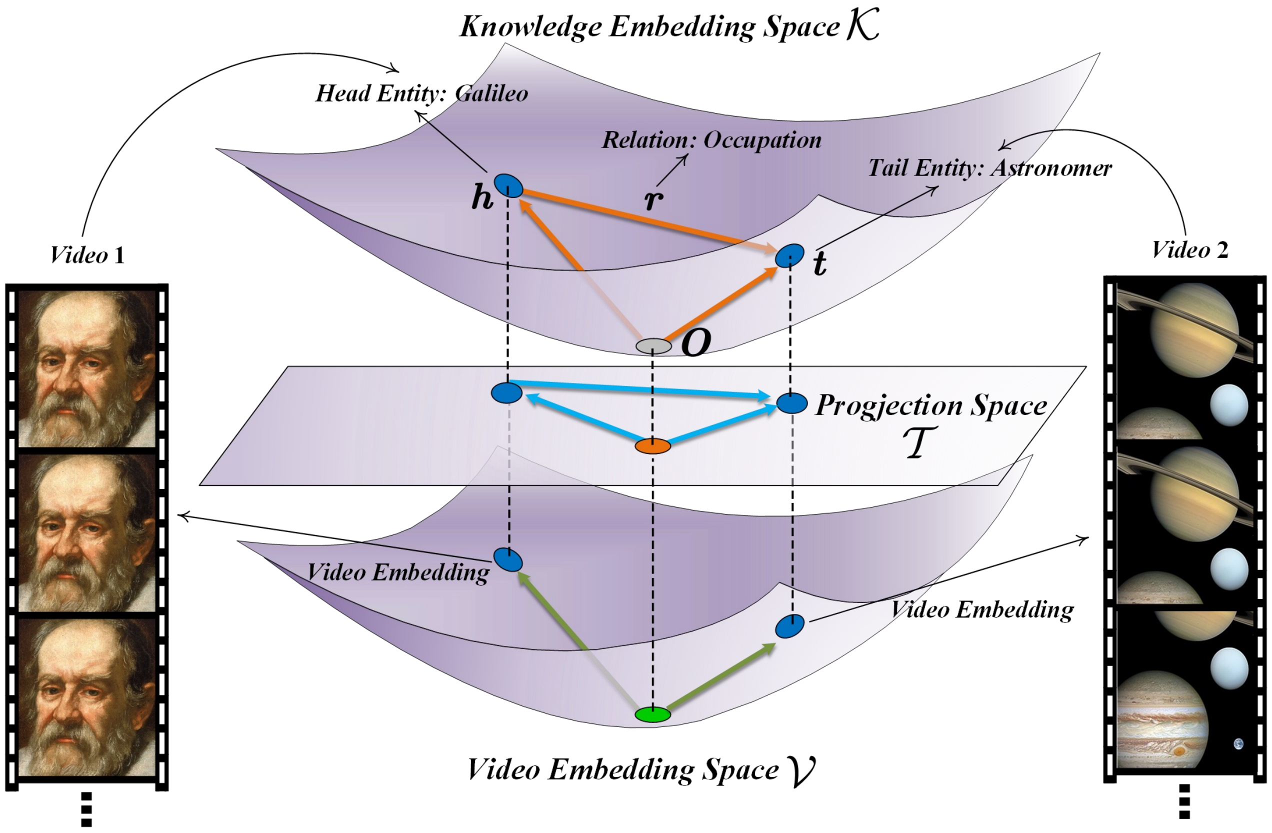


# Motivation



## ➢ Exiting Methods

- KGE: TransE[1], TransH[2], etc
  Only focus on low-dimensional space
- KGE + Language Model: K-BERT[3]
  Only focus on text modality
- KG + Video: ACAR-Net[4]
  Only focus on human activity recognition




## ➢ Three Challenges

- Form a video-based multi-modal knowledge graph dataset
- An effective embedding representation of video
- The heterogeneity issue of video and KG triplet

[1] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, *26*.
[2] Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014, June). Knowledge graph embedding by translating on hyperplanes. In Proceedings of the AAAI conference on artificial intelligence (Vol. 28, No. 1).
[3] Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., & Wang, P. (2020, April). K-bert: Enabling language representation with knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 03, pp. 2901-2908)
[4] Pan, J., Chen, S., Shou, M. Z., Liu, Y., Shao, J., & Li, H. (2021). Actor-context-actor relation network for spatio-temporal action localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 464-474

# Dataset: Video Knowledge Graph Dataset & CN-DBpedia



- **Company-400M**
  *427,24 tags*
  *47,134,152 videos*
- **Company-5M**
  *248,324 entities*
  *832,577 triplets*
  *5,150 relations*
  *84,838 tags*
- **CN-Dbpedia[5] sub**
  *101,002 entities*
  *465,714 triplets*
  *4,987 relations*

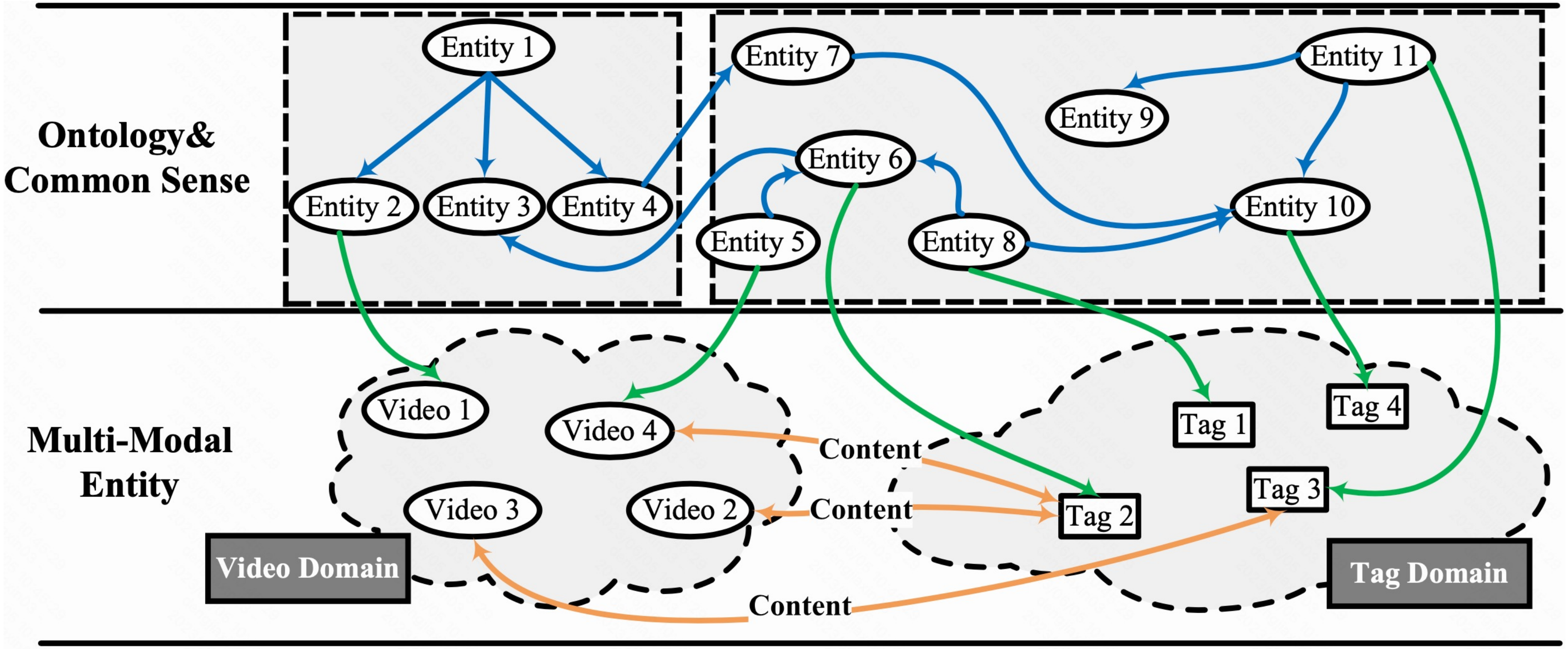



**Table 1: The meta information of the related dataset. $\mathcal{V}$, $\mathcal{A}$ and $\mathcal{T}$ represent video, audio and text respectively.**

| Dataset | Entities | Triplets | Relations | Tags | Modalities | Videos |
|---|---|---|---|---|---|---|
| Company-400M | - | - | - | 427,249 | $\{\mathcal{V},\mathcal{A},\mathcal{T}\}$ | 47,134,152 |
| Company-5M | 248,324 | 832,577 | 5,150 | 84,838 | $\{\mathcal{V},\mathcal{A},\mathcal{T}\}$ | 5,714,531 |
| CN-DBpedia sub | 101,002 | 465,714 | 4,987 | - | $\{\mathcal{T}\}$ | - |

[5] Xu, B., Xu, Y., Liang, J., Xie, C., Liang, B., Cui, W., & Xiao, Y. (2017, June). CN-DBpedia: A never-ending Chinese knowledge extraction system. In Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part II (pp. 428-438). Cham: Springer International Publishing.

# Proposed Method

$$L_{KG} = -\log\sigma(\gamma - d(\boldsymbol{h}+\boldsymbol{r},\boldsymbol{t})) - \sum_{i=1}^{n}\frac{1}{n}\log\sigma(d(\boldsymbol{h}+\boldsymbol{r},\boldsymbol{t}_i') - \gamma)$$

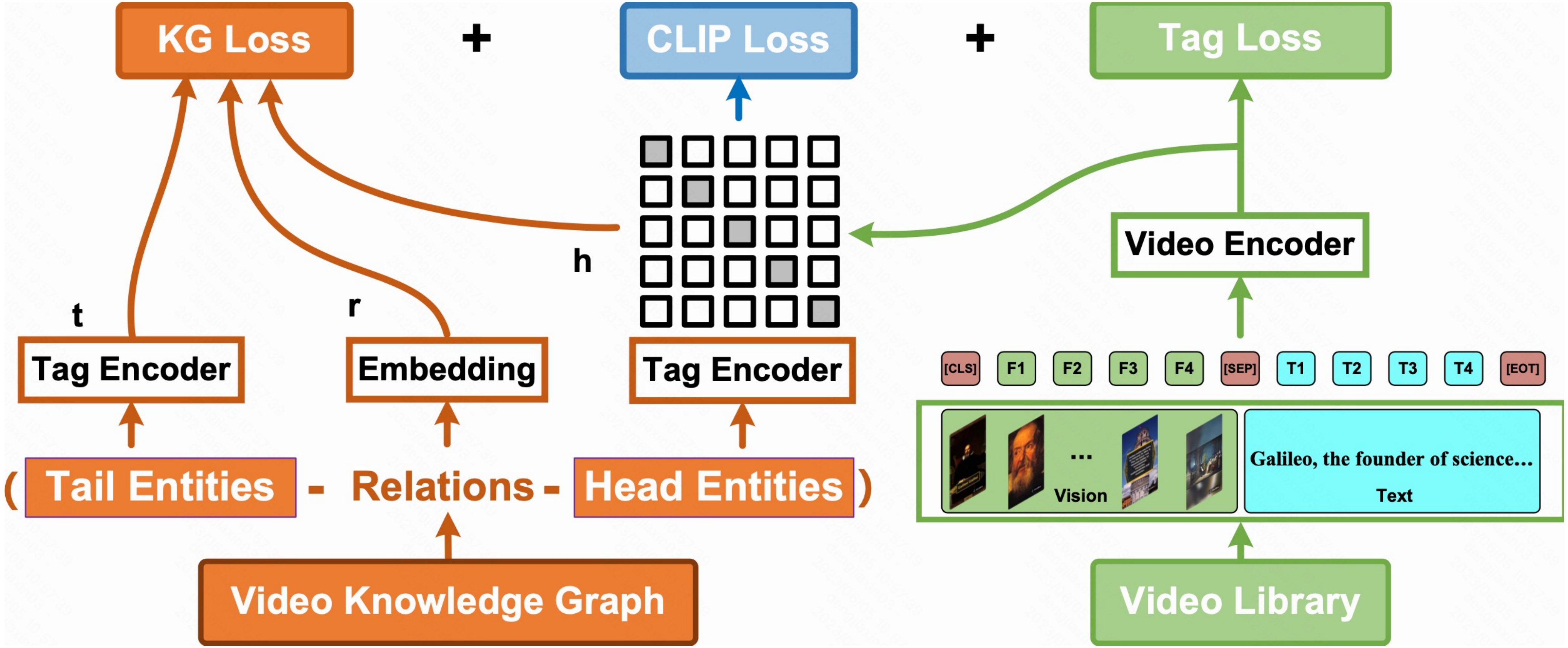
$$L_{CLIP} = \frac{1}{B}\sum_{i}^{B} -\log\frac{\exp(z_V^{(i)}\cdot z_T^{(i)}/\tau)}{\sum_{i=1}^{B}\exp(z_V^{(i)}\cdot z_T^{(j)}/\tau)} + \frac{1}{B}\sum_{i}^{B} -\log\frac{\exp(z_V^{(i)}\cdot z_T^{(i)}/\tau)}{\sum_{j=1}^{B}\exp(z_V^{(j)}\cdot z_T^{(i)}/\tau)}$$

$$L_{TAG} = -\sum_{i=1}^{T} y_i \log(s_i)$$

Stage 1

Stage 2

Stage 3



# Experiment Detail



## ➢ Evaluation Tasks

- Tag-to-Video(TV)
- Vide-to-Tag (VT)
- Tag-Relation-Tag (TRT)
- Video-Relation-Tag (VRT)
- Video-Relation-Video (VRV)

## ➢ Baselines

- TransE (Bordes et al., NIPS 2013)
- TransH (Wang et al., AAAI 2014)
- TransR (Lin et al., AAAI 2015)
- CLIP (Radford et al., ICML 2021)
- CLIP+TransE
- CLIP+TransH
- CLIP+TransR
- Ours

**Table 2: The baselines and variants of our method. $\mathcal{L}_{\mathrm{KG}}$ represents the corresponding KGE loss for TransE, TransH or TransR.**

| Baseline | VRV | VRT | TRT | VT | TV | $\mathcal{L}_{\mathrm{TAG}}$ | $\mathcal{L}_{\mathrm{CLIP}}$ | $\mathcal{L}_{\mathrm{KG}}$ |
|---|---|---|---|---|---|---|---|---|
| **TransE** | - | - | √ | - | - | - | - | √ |
| **TransH** | - | - | √ | - | - | - | - | √ |
| **TransR** | - | - | √ | - | - | - | - | √ |
| **CLIP** | - | - | - | √ | √ | √ | √ | - |
| **CLIP+TransE** | √ | √ | √ | √ | √ | √ | √ | √ |
| **CLIP+TransH** | √ | √ | √ | √ | √ | √ | √ | √ |
| **CLIP+TransR** | √ | √ | √ | √ | √ | √ | √ | √ |
| **Ours** | √ | √ | √ | √ | √ | √ | √ | √ |

## ➢ Evaluation Metrics

- Mean Rank (MR)
- Hit@n

# Experiment Result



## Content based Retrieval Task Performance

- The participation of knowledge graph embedding benefits the content retrieval task.

**Table 3: The performance comparison of VT and TV retrieval task.**

| Method | VT | | | | TV | | | |
|---|---|---|---|---|---|---|---|---|
| | MR | HITS@1 | HITS@3 | HITS@10 | MR | HITS@1 | HITS@3 | HITS@10 |
| CLIP | 14515.2419 | 0.0885 | 0.1487 | 0.2252 | 12038.8518 | 0.1143 | 0.1864 | 0.2660 |
| Ours | **10622.3440** | **0.1241** | **0.2186** | **0.3438** | **9030.5341** | **0.2786** | **0.3907** | **0.4759** |

## VRV and VRT Inference Task Performance

- Two-stage methods lack the synthetic integration of multi-modality entities and knowledge graph embeddings.

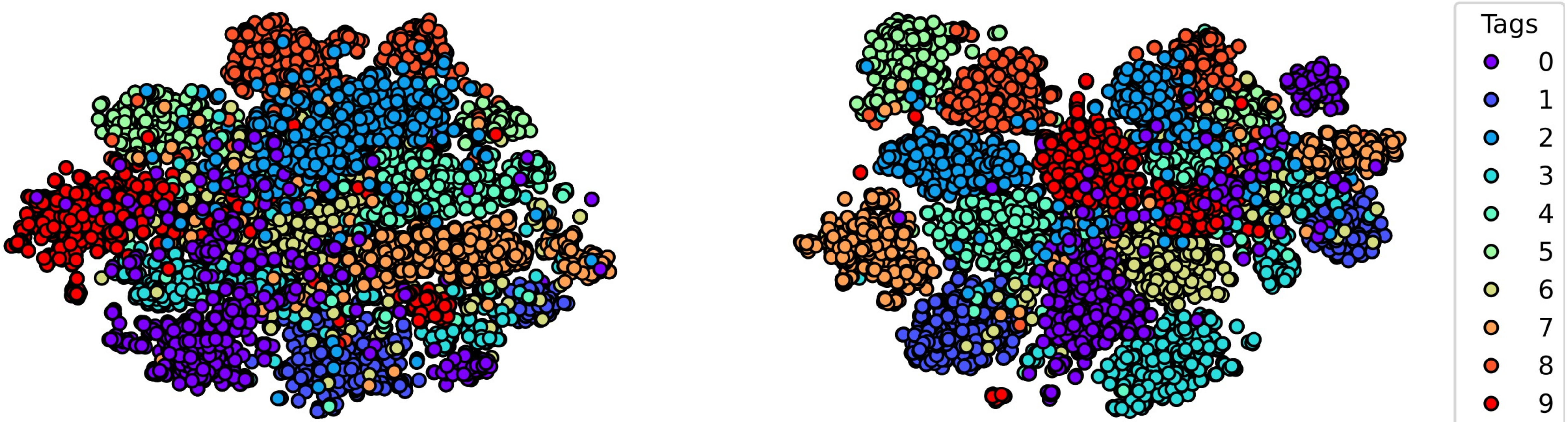
**Table 5: The performance comparison of VRV and VRT inference task.**

| Method | VRV | | | | VRT | | | |
|---|---|---|---|---|---|---|---|---|
| | MR | HITS@1 | HITS@3 | HITS@10 | MR | HITS@1 | HITS@3 | HITS@10 |
| CLIP+TransE | 23356.6640 | 0.0340 | 0.0618 | 0.0961 | 52.3991 | 0.0508 | 0.1019 | 0.3981 |
| CLIP+TransH | 23168.7382 | 0.0368 | 0.0674 | 0.1063 | 34.7941 | 0.0498 | 0.1198 | 0.4506 |
| CLIP+TransR | 27608.6244 | 0.0475 | 0.0884 | 0.1396 | 25.5660 | 0.0508 | 0.2152 | 0.5869 |
| Ours | **8357.8196** | **0.2759** | **0.3977** | **0.5632** | **13.4505** | **0.1144** | **0.4308** | **0.7642** |

# Experiment Result

## ➢ Feature Visualization

- KGE space helps the embedding of tag and video cluster better.
- KG knowledge benefit the content-based retrieval task.



# Experiment Result



## ➢ Contributions

- To the best of our knowledge, we first define a novel formulation of the Video-Relation-Video and Video-Relation-Tag inference tasks.
- We propose and form a large scale heterogeneous video knowledge graph dataset which is capable of conducting Video-Relation-Video and Video-Relation-Tag inference tasks.
- We propose a transformer architecture for multi-modal video understanding and knowledge graph embedding integration.
- Extensive experiments indicate that our method achieves the *state-of-the-art* performance on video inference tasks and it also brings improvement on content-based video retrieval tasks.

## ➢ Future Work

- Conduct more experiments on public multi-modal knowledge graph datasets such as FB15K237[6].
- Explore more advanced approaches to integrate video understanding and KG semantic space.

**[6] Toutanova, K., & Chen, D. (2015, July). Observed versus latent features for knowledge base and text inference. In Proceedings of the 3rd workshop on continuous vector space models and their compositionality (pp. 57-66).**