ORANGE: Text-video Retrieval via Watch-time-aware Heterogeneous Graph Contrastive Learning

Yucheng Lin [1*] **Yaning Chang** [1*] **Tim Chang** [2*] **Jianqiang Ma** [3] **Donghui Li** [1]
Ting Peng [1] **Zang Li** [1] **Zhiyi Zhou** [1] **Feng Wang** [1]
PCG, Tencent, China
[1]{ycl,carloschang,dhlli,penneypeng,gavinzli,liuxizhou,feynmanwang}@tencent.com
[2]timchang2022@163.com
[3]majianchiang@gmail.com

Abstract

With the explosive growth of short-video data on industrial video-sharing platforms such as TikTok and YouTube, text-video retrieval techniques have become increasingly important. Most existing works for text-video retrieval focus on designing informative representation learning methods and delicate matching mechanisms, which leverage the content information of queries and videos themselves (*i.e.*, textual information of queries and multimodal information of videos). However, real-world scenarios often involve brief, ambiguous queries and low-quality videos, making content-based retrieval less effective. In order to accommodate various search requirements and enhance user satisfaction, this study introduces a novel Text-vide**o** **R**etrieval method vi**a** Watch-time-aware Heteroge**n**eous **G**raph Contrastiv**e** Learning (termed ORANGE). This approach aims to learn informative embeddings for queries and videos by leveraging both content information and the abundant relational information present in video-search scenarios. Specifically, we first construct a heterogeneous information graph where nodes represent domain objects (*e.g.*, query, video, tag) and edges represent rich relations among these objects. Afterwards, a meta-path-guided heterogeneous graph attention encoder with the awareness of video watch time is devised to encode various semantic aspects of query and video nodes. To train our model, we introduce a meta-path-wise contrastive learning paradigm that facilitates capturing dependencies across multiple semantic relations, thereby enhancing the obtained embeddings. Finally, when deployed online, for new queries non-existent in the constructed graph, a bert-based query encoder distilled from our ORANGE is employed. Offline experiments conducted on a real-world dataset demonstrate the effectiveness of our ORANGE. Moreover, it has been implemented in the matching stage of an industrial online video-search service, where it exhibited statistically significant improvements over the online baseline in an A/B test.

1 Introduction

With the exponential proliferation of short-video data on the internet, text-video retrieval (Wang et al., 2022; Liu et al., 2022; Bain et al., 2021; Gabeur et al., 2020; Gorti et al., 2022; Luo et al., 2022) has gained increasing attention from both industrial and academic communities, and become a crucial feature of industrial video-sharing platforms (*e.g.*, TikTok, Likee, and YouTube). The goal of text-video search is to retrieve the most user-satisfactory videos given a text query. Towards this end, great efforts have been devoted to carefully designing informative representation learning methods (Zhao et al., 2022; Bain et al., 2021; Xiao et al., 2022; Arnab et al., 2021; Vaswani et al., 2017; Wang et al., 2022; Luo et al., 2022) and delicate text-video matching mechanisms using single-stream or dual-stream architectures (Gorti et al., 2022; Min et al., 2022; Zhu and Yang, 2020; Lei et al., 2021; Liu et al., 2021; Wang et al., 2022; Zhao et al., 2022; Luo et al., 2022). For example, CLIPBERT (Lei et al., 2021) jointly embeds text-video pairs through a BERT-like (Devlin et al., 2018) single-stream encoder for early cross-modal fusion and directly produces similarity between them. CLIP4Clip (Luo et al., 2022) introduces a dual-stream encoder consisting of a transformer encoder (Vaswani et al., 2017) for texts and a space-time transformer encoder for videos. The obtained representations of texts and videos are then mapped to a common space, where the text-video similarity is measured via the dot product.

The above text-video retrieval methods mainly focus on utilizing the content information of queries and videos themselves, *i.e.*, textual information from queries and multimodal information from videos, including video titles, video frames, audio, etc. Despite their effectiveness, these meth-

[*] Equal contribution

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 275–283
December 6-10, 2023 ©2023 Association for Computational Linguistics

ods face two challenges when applied to industrial video-search scenarios. First, unlike academic datasets (Chen et al., 2015; Miech et al., 2019) used in previous works, real-world query texts tend to be shorter and ambiguous, while user-generated(or uploaded) videos may exhibit low quality. Consequently, text-video relevance matching based solely on content information is insufficient for accurately capturing search intents. Secondly, for certain queries, the videos retrieved by these content-based methods are likely to be highly relevant, making it difficult to discern the most user-satisfactory videos from semantically similar candidates. To address the aforementioned challenges, we propose a novel Text-vide**o** **R**etrieval method vi**a** Watch-time-aware Heteroge**n**eous **G**raph Contrastiv**e** Learning (referred to as ORANGE) in this paper. ORANGE aims to learn discriminative embeddings for queries and videos, taking into account not only content information but also the abundant relational information present in video-search scenarios. Concretely, based on the text-video search log and domain knowledge, we first construct a heterogeneous information graph(HIG) (as shown in Fig. 1(a)), where nodes represent video-retrieval domain objects (*e.g.*, query, video, video tag), and edges represent rich relations among these objects. For instance, a query-video edge describes the relationship where a video is viewed given a query, while a query-query edge describes the rewriting relation between a pair of queries. In order to fully utilize the rich heterogeneous information and thereby learn informative representations, a meta-path-guided heterogeneous graph attention encoder (HAN) (Wang et al., 2019) with the awareness of the video watch time is devised to encode various semantic aspects of query and video nodes. The vanilla graph attention mechanisms (Wang et al., 2019; Veličković et al., 2017a) are sometimes ineffective as the attention weights are learned implicitly without the guidance of explicit semantics (Jain and Wallace, 2019). In contrast, our watch-time-aware encoder enhances the quality of attention weights by explicitly incorporating the video's watch-time information, which is a crucial indicator of user satisfaction. To train ORANGE, we cast the text-video matching problem as a link prediction task of HIG. Simultaneously, we employ an auxiliary learning task based on our proposed meta-path-wise contrastive learning paradigm, which helps the model capture cross-type semantic dependencies and improve the quality of embeddings. Lastly, when deploying our model online, we adopt a BERT-based query encoder distilled (Hinton et al., 2015) from ORANGE, enabling on-the-fly inference of query embeddings to support previously unseen queries, *i.e.*, new queries non-existent in our HIG. To the best of our knowledge, we are the first on utilizing both rich relational information and content information for text-video retrieval in real-world scenarios.

In a nutshell, this work makes the following contributions:

- We build a heterogeneous information graph (HIG) to comprehensively integrate content information and rich relational information existing in video-search scenarios, which is then encoded by our meta-path-guided heterogeneous graph neural network.

- Our newly-devised watch-time-aware encoder can improve the vanilla graph attention mechanism by explicitly injecting the video's watch-time information which is an important indicator of user satisfaction.

- To ensure robust learning, we leverage a meta-path-wise contrastive learning strategy to capture dependencies of the cross-type semantic relations of HIG and then enhance the obtained representations.

- Considering the online deployment, we also further propose a graph distillation strategy that allows our distilled query encoder to deal with unseen queries.

2 Methodology

In this section, we describe the details of our ORANGE approach. The overall workflow of our model is shown in Fig. 1 and the detailed notations used in this paper are summarized in Table 4 in Appendix.

2.1 HIG Construction

To obtain informative representation, a heterogeneous information graph (HIG) is first built to comprehensively integrate content information and rich relational information existing in industrial scenarios. Formally, we define the HIG for our video search scenarios as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X})$, where $\mathcal{V}$, $\mathcal{E}$, and $\mathcal{X}$ denote the sets of nodes, edges, and attributed features, respectively. These are

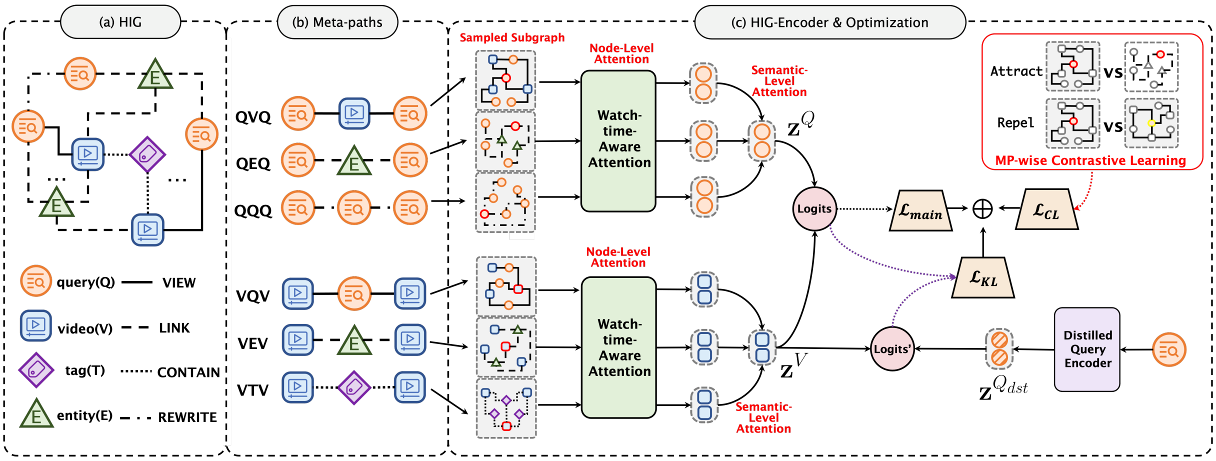


Figure 1: Workflow of the proposed ORANGE. It consists of three key components: (a) HIG construction, (b) Semantic meta-path design, and (c) Encoder and Optimization.

associated with a node type mapping function $\phi : \mathcal{V} \rightarrow \mathcal{R}^v$ and an edge type mapping function $\varphi : \mathcal{E} \rightarrow \mathcal{R}^e$, where $\mathcal{R}^v$ and $\mathcal{R}^e$ denote the types of nodes and edges respectively. An example of the HIG is illustrated in Fig. 1(a) and Fig. 1(b). There are four different types of nodes (*i.e.*, query(**Q**), video(**V**), entity(**E**) and tag(**T**) and four different types of edges(*i.e.*, **VIEW**, **LINK**, **CONTAIN** and **REWRITE**).

In a heterogeneous graph, two objects can be connected via different semantic paths, which are referred to as meta-paths (Wang et al., 2019)(denoted as $\Phi \in \mathcal{M}$, where $\mathcal{M}$ represents the predefined set of meta-paths. To extract the semantic relations among diverse types of nodes, in this work, we design three distinct types of meta-paths(*i.e.*, **QVQ**, **QEQ**, and **QQQ**) for query nodes and another three distinct meta-paths (*i.e.*, **VQV**, **VTV**, and **VEV**) for video nodes. These different meta-paths can reveal different semantics and complement each other. For instance, **QEQ** denotes two queries linked by the same **entities** (*e.g.*, **actors, movies, songs, etc., included in the queries or videos**, as illustrated in Appendix B), which may indicate two queries share the same search intent. **QVQ** implies that a single video is viewed under two different queries, hinting at a semantic relevance between the two queries. **QQQ** represents consecutive queries within a search session, potentially indicating rewriting behaviors when the retrieved results are unsatisfactory. Note that we emphasize the generality of our graph construction approach, which is applicable to most real-world search scenarios.

2.2 HIG Encoder

To encode the rich heterogeneous information of HIG, we devise a meta-path-guided heterogeneous graph attention network with the awareness of the video watch time (abbreviated as HIG encoder). Similar to the work in HAN (Wang et al., 2019), we employ a two-stage attention mechanism to encode node representations, namely node-level attention and semantic-level attention, as shown in Fig. 1(c). For each node i of the HIG and a specified meta-path Φ, the **node-level** attention aims to learn the importance of meta-path-based neighbors $j \in \mathcal{N}_i^{\Phi}$, where $\mathcal{N}_i^{\Phi}$ denotes the meta-path-based neighbors of node i. The attention coefficients e_{ij}^{Φ}, indicating the importance of node i to node j, are computed as follows:

$$e_{ij}^{\Phi} = \sigma\left(\mathbf{a}_{\Phi}^{\mathrm{T}} \cdot [W_{\Phi}^{l}\mathbf{h}_i^{l,\Phi} \| W_{\Phi}^{l}\mathbf{h}_j^{l,\Phi}]\right) + \psi_{ij},$$

$$\psi_{ij} = \begin{cases} \lambda \cdot t_v & if\ \varphi_{ij} = \textbf{VIEW}, \\ 0 & otherwise. \end{cases} \tag{1}$$

Here $\mathbf{h}_i^{l,\Phi} \in \mathbb{R}^d$ denotes the node representation in the l-th layer of our HIG encoder, where l=0,1,2. $\mathbf{h}_i^{0,\Phi}$ is the projected representation of the attributed feature $x_i \in \mathcal{X}$ for node i. $\mathbf{a}_{\Phi}$ denotes the **node-level** attention vector for meta-path Φ, and W_{Φ}^{l} is the meta-path-specific transformation matrix. $\|$ denotes the concatenation operator and $\sigma(\cdot)$ denotes the non-linear activation function. It is noteworthy that, in contrast to traditional graph attention mechanisms (Wang et al., 2019) where attention weights are determined implicitly without explicit semantics, we explicitly incorporate prior watch-time information into the original attention calculation in order to make our encoder aware of the watch-time information, which is a key factor in user satisfaction. Since videos with longer duration tend to have longer watch time, to alleviate the video duration bias, we follow the method from (Zheng et al., 2022) to transform the original video watch-time into an unbiased watch-time score $t_v \in \mathbb{R}$. $\lambda \in \mathbb{R}$ is a learnable parameter. Note that the watch-time aware attention only applies to the **VIEW** edge type. After obtaining the coefficients between node pairs based on the meta-path Φ, we normalize them via the softmax function:

$$\alpha_{ij}^{\Phi} = \frac{\exp(e_{ij}^{\Phi})}{\sum_{k \in \mathcal{N}_i^{\Phi}} \exp(e_{ik}^{\Phi})}, \tag{2}$$

and then compute the output representation of node i corresponding to the meta-path Φ as follows:

$$\mathbf{h}_i^{l+1,\Phi} = \sigma(\sum_{j \in \mathcal{N}_i^{\Phi}} \alpha_{ij}^{\Phi} W_{\Phi}^{l} \mathbf{h}_j^{l,\Phi}). \tag{3}$$

After we obtain node-level representations corresponding to different semantic meta-paths, our **semantic-level** attention is used to get the final informative representation $\mathbf{z}_i$ of node i as follows:

$$\begin{aligned} \omega_i^{\Phi} &= \mathbf{MLP}(\mathbf{h}_i^{\Phi}), \Phi \in \mathcal{M}, \\ \beta_i^{\Phi} &= \exp(\omega_i^{\Phi}) / \sum_{\Phi' \in \mathcal{M}} \exp(\omega_i^{\Phi'}), \\ \mathbf{z}_i &= \sum_{\Phi \in \mathcal{M}} \beta_i^{\Phi} \mathbf{h}_i^{l=2,\Phi} \end{aligned} \tag{4}$$

Here, we can obtain query representation $\mathbf{z}^Q$ and video representation $\mathbf{z}^V$ when $\mathcal{M}^Q$={**QVQ**, **QEQ**, **QQQ**} and $\mathcal{M}^V$={**VQV**, **VEV**, **VTV**} respectively.

2.3 Model Optimization

As for optimization, our model loss(as shown in Fig. 1(c)), is comprised of three parts, namely **main loss**, **contrastive loss** and **graph distillation loss**, respectively.

2.3.1 Main Loss

We cast our problem of learning $\mathbf{z}^Q$ and $\mathbf{z}^V$ as a link prediction task of HIG. Formally, let $\mathcal{S} = \{m_p, n_p\}_{p=1}^{|\mathcal{S}|}$ be a training batch of query and viewed videos pairs sampled from the search logs. Given the node representation $\mathbf{z}_{m_p}^Q$ for the query m_p and the node representation $\mathbf{z}_{n_p}^V$ for the engaged video n_p, we optimize them by minimizing the following in-batch loss:

$$\mathcal{L}_{main} = -\frac{1}{|\mathcal{S}|} \sum_{p=1}^{|\mathcal{S}|} \log \frac{\exp\left(\langle \mathbf{z}_{m_p}^Q, \mathbf{z}_{n_p}^V \rangle / \tau_1\right)}{\sum_{p'=1}^{|\mathcal{S}|} \exp\left(\langle \mathbf{z}_{m_p}^Q, \mathbf{z}_{n_{p'}}^V \rangle / \tau_1\right)}, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity and τ_1 is a learnable temperature parameter.

2.3.2 Contrastive Loss

Different meta-paths can reveal different semantics. We propose a meta-path-wise contrastive learning loss to capture the complex dependencies across different types of meta-paths. Specifically, taking the query representation for example, given a pair of node embedding $(\mathbf{h}_{m_p}^{\Phi}, \mathbf{h}_{m_{p'}}^{\Phi'})$ corresponding to meta-paths Φ and Φ' respectively($\Phi \neq \Phi'$), it can be regarded as a positive pair when $p = p'$ and negative when $p \neq p'$. Our meta-path-wise contrastive loss is formulated as follows:

$$\begin{aligned} \ell_{cl,p}^{Q}(\Phi, \Phi') &= -\log \frac{\exp(\langle \mathbf{h}_{m_p}^{\Phi}, \mathbf{h}_{m_p}^{\Phi'} \rangle / \tau_2)}{\sum_{p'=1}^{|\mathcal{S}|} \exp(\langle \mathbf{h}_{m_p}^{\Phi}, \mathbf{h}_{m'_p}^{\Phi'} \rangle / \tau_2)}, \\ \mathcal{L}_{cl}^{Q} &= \frac{1}{|\mathcal{S}|} \sum_{p=1}^{|\mathcal{S}|} \sum_{\Phi, \Phi' \in \mathcal{M}^Q} \ell_{cl,p}^{Q}(\Phi, \Phi') \end{aligned} \tag{6}$$

where τ_2 is the learnable temperature parameter. Similarly, we can obtain the meta-path-wise contrastive loss for video nodes, i.e., $\mathcal{L}_{cl}^V$.

2.3.3 Graph Distillation Loss

For online deployment, we devise a graph distillation strategy that allows our distilled query encoder to handle unseen queries. Specifically, our distilled query encoder is a 4-layer BERT and the distilled query representation, denoted as $\mathbf{z}^{Q_{dst}} \in \mathbb{R}^d$, can be optimized by the following loss:

$$\begin{aligned} P(\mathbf{z}_{n_p}^V | \mathbf{z}_{m_p}^Q) &= \frac{\exp\left(\langle \mathbf{z}_{m_p}^Q, \mathbf{z}_{n_p}^V \rangle\right)}{\sum_{p'=1}^{|\mathcal{S}|} \exp\left(\langle \mathbf{z}_{m_p}^Q, \mathbf{z}_{n_{p'}}^V \rangle\right)}, \\ P(\mathbf{z}_{n_p}^V | \mathbf{z}_{m_p}^{Q_{dst}}) &= \frac{\exp\left(\langle \mathbf{z}_{m_p}^{Q_{dst}}, \mathbf{z}_{n_p}^V \rangle\right)}{\sum_{p'=1}^{|\mathcal{S}|} \exp\left(\langle \mathbf{z}_{m_p}^{Q_{dst}}, \mathbf{z}_{n_{p'}}^V \rangle\right)}, \\ \mathcal{L}_{kl} &= \sum_{p=1}^{|\mathcal{S}|} \mathbf{KLD}\left(P(\mathbf{z}_{n_p}^V | \mathbf{z}_{m_p}^Q), P(\mathbf{z}_{n_p}^V | \mathbf{z}_{m_p}^{Q_{dst}})\right) \end{aligned} \tag{7}$$

where **KLD** denotes the KL-divergence.

2.3.4 Complexity

The overall time complexity is $\mathcal{O}(|\mathcal{V}|d^2 + |\mathcal{E}|d)$, where d denotes the dimension of node embeddings. The parallelization of the proposed model is easily achievable, since both the node-level and semantic-level attention can be concurrently processed across node pairs and meta-paths, respectively.

3 Experiments

3.1 Datasets

Training Dataset. Since the existing public dataset (Luo et al., 2022) for the text-video retrieval tasks lack the watch-time and relational information that are widely present in the real-world video-search scenario, we utilize an industrial dataset derived from the search logs of Tencent Video, a prominent Chinese video streaming platform. This data (as illustrated in Appendix B), collected from July 1st to August 30th, 2022, is used to construct our HIG, comprising approximately 26 million nodes and 215 million edges. The detailed statistics of our constructed HIG are shown in Table 1.

Node	Edge	Meta-paths
# query(Q): 321K # video(V): 25M # entity(E): 109K # tag(T): 80K	# Q-V: 34M # Q-E: 239K # Q-Q: 1M # V-E: 26M # V-T: 47M	QVQ QEQ QQQ VQV VEV VTV

Table 1: Statistics of the constructed HIG.

Evaluation Dataset. Our evaluation dataset, collected from logs on August 31st, is divided into two subsets. Subset-1 includes queries in the constructed HIG, while Subset-2 comprises unseen, typically long-tail, queries. Subset-1 holds around 60,000 positive pairs (Query-Video) from nearly 7,000 queries and 55,000 videos. Subset-2 has about 3,000 positive pairs, with roughly 1,900 queries and 2,900 videos. For each positive pair, we randomly select 10 negative videos for evaluation.

3.2 Comparison Methods and Metric

To verify the effectiveness of our proposed method, we choose the following comparison methods for evaluation.

SBERT (Reimers and Gurevych, 2019): it is a text-based model that uses a BERT-like dual tower network to derive semantically meaningful embeddings. In this setting, we only consider the textual information. Both the query and the textual information of videos are encoded using a shared BERT encoder. The video representation is a concatenation of video title, video tags and entities.

CLIP4Clip (Luo et al., 2022): it is a multimodal-based model including a text encoder and a space-time encoder that extract representations of texts and videos respectively.

GAT (Veličković et al., 2017b): it is a widely-used graph neural network which performs attention operation on graphs. In this setting, we only consider the query-video interaction to construct a graph where a 2-layer GAT is used.

All baselines are evaluated on the metric $Recall@K$ ($R@K$, K=10, 50, 100) which measures how many correct videos are recalled within the top K results, and $AWT@100$ which evaluates the Average Watch Time (**seconds**) of the top 100 candidates. We exclude CLIPBERT from our comparison as its single-stream encoder is not applicable to the matching stage of online video-search services.

3.3 Implementation Details

We implement all models using Tensorflow and Adam (Nothaft et al., 2015) optimizer with a fixed learning rate of $7e^{-4}$. We train our model for $20,000$ steps with a batch size of $4,096$. Additionally, we employ dropout with a drop rate of 0.1 to alleviate the overfitting issue. The temperatures τ_1 and τ_2 are both initially set to 0.05 and the output embedding dimension is 64 for all models. The maximum number of each-hop neighbors for GAT and ORANGE is set to 10. As for the attributed features of GAT and ORANGE, we employ a pre-trained BERT encoder from CLIP4Clip to encode textural information of queries, titles, entities and tags and a pre-trained video encoder from CLIP4Clip to encode visual information. To get the full multi-modal features of videos, we concatenate the textural and visual features. As for SBERT and CLIP4Clip, we employ the default setting as described in their original papers.

3.4 Evaluations

3.4.1 Offline Evaluation

As shown in Table 2, our proposed ORANGE significantly outperforms baseline models on both the R@K and AWT@100 metrics. As expected, CLIP4Clip performs better than the text-based model SBERT in terms of R@K, which indicates

Models	R@10	R@50	R@100	AWT@100
SBERT	0.429	0.640	0.712	19.9
CLIP4Clip	0.450	0.673	0.744	19.7
GAT	0.576	0.809	0.870	21.3
w/o QEQ	0.621	0.859	0.909	25.0
w/o VEV	0.622	0.862	0.913	25.1
w/o VTV	0.616	0.862	0.904	23.9
w/o QQQ	0.614	0.854	0.899	24.1
w/o CL	0.615	0.853	0.891	23.4
w/o time-aware	0.614	0.865	0.889	22.5
ORANGE$_{dst}$	0.567	0.791	0.850	21.7
ORANGE	**0.627**	**0.868**	**0.919**	**25.2**

Table 2: Offline comparison on Subset-1.

that using additional visual information can benefit text-video matching. GAT, by leveraging simple structural information and attributed information, considerably outperforms both SBERT and CLIP4Clip, indicating the importance of behavior information in text-video matching. However, GAT still fails to take enough heterogeneity and video watch time into consideration. Our model that fully utilizes both content information and rich relational information beats GAT by 8.9%, 7.3% and 5.6% on the R@K metric (K=10,50,100) and by 18.3% on the AWT@100 metric.

We assess the performance of the distilled ORANGE model. As shown in Table 2, despite the model capacity loss between the distilled and the full version, the distilled ORANGE still shows comparative performance to GAT and significantly outperforms the two content-based baselines, SBERT and CLIP4Clip. Meanwhile, we also evaluate the distilled version on Subset-2 (unseen queries) where it still beats SBERT and CLIP4Clip slightly.

Models	R@10	R@50	R@100	AWT@100
SBERT	0.588	0.767	0.810	3.597
CLIP4Clip	0.638	**0.787**	**0.829**	3.760
GAT	-	-	-	-
ORANGE$_{dst}$	**0.655**	**0.787**	0.826	**3.937**

Table 3: Offline results on Subset-2(unseen queries).

3.4.2 Ablation Studies

In the ablation experiment, we evaluate the contribution of each component to the improvement of model performance.

Effect of semantic meta-paths. Different meta-paths can represent different semantics. We evaluate the effect of various meta-paths by removing the corresponding nodes and edges in turn. There are three sub-groups, namely, entity-related, tag-related and rewriting, respectively. As shown in Table 2, the performances consistently decline compared to the full model, illustrating that rich relational information can assist in obtaining informative representations.

Effect of meta-path-wise contrastive learning. To investigate whether the meta-path-wise contrastive learning strategy benefits the model training, we train another ablation model without using contrastive loss (w/o CL). According to the table 2, the R@10 drops from 0.627 to 0.615 when ablating the contrastive loss. It suggests that our contrastive loss can help enhance the quality of representations by maximizing mutual information between different semantic meta-path views of the same nodes.

Effect of watch-time-aware attention. Without utilizing the watch-time-aware attention encoder, the performance of AWT@100 drastically drops from 25.2 to 22.5, which demonstrates that our method of explicitly injecting semantic information (*e.g.*, video watch time) into the vanilla attention calculation enhances the informativeness of the obtained embeddings.

3.4.3 Deployment & Online A/B Test

We conducted our online experiment on the matching stage of our online video-search service. The current online video-search engine is a highly optimized system with multiple retrieval routes to provide candidates, which are subsequently processed by ranking modules. Specifically, we utilize the distilled ORANGE for unseen nodes and the full version for nodes within the pre-constructed graph. The online control group contains the matching methods mainly based-on textual and visual information such as used in SBERT (Reimers and Gurevych, 2019) and CLIP4Clip (Luo et al., 2022). Both the constructed HIG and our trained ORANGE model are incrementally updated on a daily basis. We have observed a statistically significant cumulative improvement by 2.08% in terms of AWT in an A/B test.

4 Conclusion

In this study, we address the text-video retrieval challenge by incorporating content and intricate relations among video-retrieval domain objects into a heterogeneous information graph. Based on this, we introduce a novel watch-time-aware encoder and a meta-path-wise contrastive learning strategy to obtain informative representations. Furthermore, to ensure our model's applicability for online deployment, we employ a BERT-based query encoder, distilled from our full model, to process previously unseen nodes. In our future work, we plan to explore efficient graph distillation strategies. Concurrently, the development of informative graph encoders that consider abundant search behavior information should also be investigated.

Limitations

The proposed method heavily relies on pre-defined meta-paths, and due to computational complexity, their lengths are all set to 3, which may limit the expressive capability of our model. To alleviate this issue, an automatic method need to be designed to identify useful meta-paths. Simultaneously, although we employ a distilled version to manage unseen queries, it still exhibits a substantial performance decline compared to the full version. To alleviate this issue, we may consider incrementally constructing the HIG on an hourly basis, rather than the current daily updates, based on newly acquired user search behavior data.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer.

Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5006–5015.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341.

Peidong Liu, Dongliang Liao, Jinpeng Wang, Yangxin Wu, Gongfu Li, Shu-Tao Xia, and Jin Xu. 2022. Multi-task ranking with user behaviors for text-video search. In *Companion Proceedings of the Web Conference 2022*, pages 126–130.

Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11915–11925.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Shaobo Min, Weijie Kong, Rong-Cheng Tu, Dihong Gong, Chengfei Cai, Wenzhe Zhao, Chenyang Liu, Sixiao Zheng, Hongfa Wang, Zhifeng Li, et al. 2022. Hunyuan_tvr for text-video retrivial. *arXiv preprint arXiv:2204.03382*.

Frank A Nothaft, Matt Massie, Timothy Danford, Zhao Zhang, Uri Laserson, Carl Yeksigian, Jey Kottalam, Arun Ahuja, Jeff Hammerbacher, Michael Linderman, Michael Franklin, Anthony D. Joseph, and David A. Patterson. 2015. Rethinking data-intensive science using scalable analytics systems. In *Proceedings of the 2015 International Conference on Management of Data (SIGMOD '15)*. ACM.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017a. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017b. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The world wide web conference*, pages 2022–2032.

Xun Wang, Bingqing Ke, Xuanping Li, Fangyu Liu, Mingyu Zhang, Xiao Liang, and Qiushi Xiao. 2022. Modality-balanced embedding for video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2578–2582.

Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 2022. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer.

Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. Centerclip: Token clustering for efficient text-video retrieval. *arXiv preprint arXiv:2205.00823*.

Yu Zheng, Chen Gao, Jingtao Ding, Lingling Yi, Depeng Jin, Yong Li, and Meng Wang. 2022. Dvr: Micro-video recommendation optimizing watch-time-gain under duration bias. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 334–345.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755.

Appendix

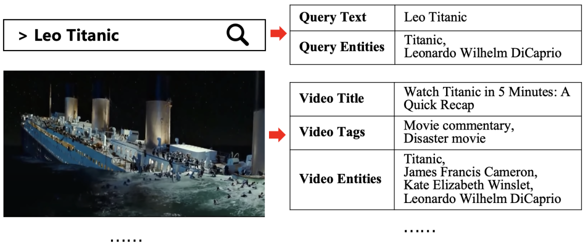
A Notations

Notation	Explanation
ϕ	Node type mapping function
φ	Edge type mapping function
Φ	Meta-path
$\mathcal{M}$	Meta-path set
$\mathcal{M}^Q$	Meta-path set for query node
$\mathcal{M}^V$	Meta-path set for video node
Q	Query node
V	Video node
T	Video tag node
E	Entity node
$\mathcal{N}^\Phi$	Neighbors per meta-path Φ
W^Φ	Projection matrix per meta-path Φ
$\mathbf{h}^\Phi$	Node representation per meta-path Φ
$\mathbf{z}$	Final aggregated node representation

Table 4: Notations and explanations.

B Data example

In this section, we show an example of our collected query-video data in Fig. 2.



Query Text	Leo Titanic
Query Entities	Titanic, Leonardo Wilhelm DiCaprio

Video Title	Watch Titanic in 5 Minutes: A Quick Recap
Video Tags	Movie commentary, Disaster movie
Video Entities	Titanic, James Francis Cameron, Kate Elizabeth Winslet, Leonardo Wilhelm DiCaprio

......

Figure 2: Illustration of our collected query-video data.