Quantum Interference-Inspired Who-What-Where Composite-Semantics Instance Search for Story Videos

Zijun Xu*
2019300003083@whu.edu.cn
Wuhan University
School of Cyber Science and Engineering
Wuhan, China

Jiahao Guo*
2018302110068@whu.edu.cn
Wuhan University
School of Computer Science
Wuhan, China

Chunjie Zhang
cjzhang@bjtu.edu.cn
Beijing Jiaotong University
School of Computer Science and Technology
Beijing, China

Zhongyuan Wang
wzyhope@163.com
Wuhan University
School of Computer Science
NERCMS & MMNCE
Wuhan, China

Chunxia Xiao
cxxiao@whu.edu.cn
Wuhan University
School of Computer Science
NERCMS & MMNCE
Wuhan, China

Chao Liang†
cliang@whu.edu.cn
Wuhan University
School of Computer Science
NERCMS & MMNCE
Wuhan, China

Abstract

The Who-What-Where (3W) composite-semantics video Instance Search (INS) task aims to find video shots about a person doing an action in a location. The state-of-the-art (SOTA) methods decompose 3W INS into three 2W INS, *i.e.*, who-what, what-where and where-who semantic correlation modeling, and directly multiply three 2W INS results to produce the final 3W INS result. Obviously, overlapping semantics exist among the above 2Ws, *e.g.*, who-what and what-where share the action component. The semantic overlap indicates that the 2Ws are mutually interdependent rather than independent. According to probability theory, the product of interdependent variables cannot be directly multiplied to obtain an accurate result, and such a direct product would yield a suboptimal outcome. This interdependence exerts diverse influences on the 3W INS results. For instance, fusing two 2W INS results "Dr. Kelleher-provide medical guidance" and "provide medical guidance-in the hospital", "provide medical guidance" is a pivotal connection, of positively enhancing the rationality of both person and location. Conversely, while both "Ross-lifts heavy objects" and "lift heavy objects-Ross" are individually coherent, combining them by overlapping the shared element "Ross" creates a conflict between the hazardous setting and strenuous labor, ultimately undermining the overall plausibility. Inspired by quantum interference theory, we propose a Quantum Interference Partial Decomposition (QIPD) method to model the diverse influences of semantic overlap from 2W to 3W INS. Specifically, QIPD incorporates two core modules, *i.e.*, semantic interference and temporal interference. The former derives the 3W amplitude by converting 2W samples into amplitudes and phases and performing interference, while the latter sets the current shot's phase as baseline, amplifying the influence of adjacent shots while attenuating distant shots. Extensive evaluations on three large-scale 3W INS datasets demonstrate that QIPD outperforms SOTA baselines.

CCS Concepts

• **Information systems → Video search**.

Keywords

Who-What-Where, instance search, quantum interference, partial decomposition

ACM Reference Format:
Zijun Xu, Jiahao Guo, Chunjie Zhang, Zhongyuan Wang, Chunxia Xiao, and Chao Liang. 2025. Quantum Interference-Inspired Who-What-Where Composite-Semantics Instance Search for Story Videos. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (ACM MM).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

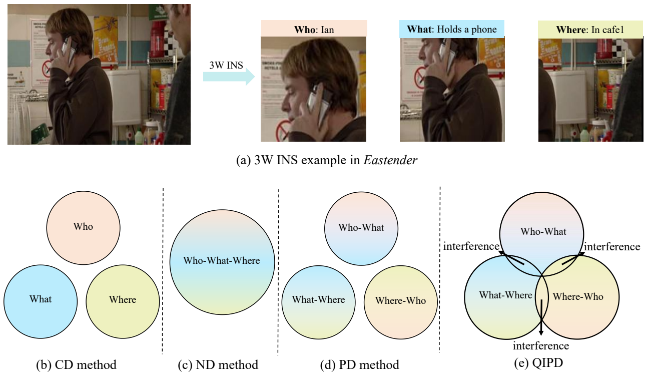
The Who-What-Where (3W) composite-semantics Video Instance Search (INS) task, also known as Person-Action-Location (PAL) INS, represents a challenging problem in the field of computer vision. This task requires retrieving video shots containing specific combinations of three semantic elements, *i.e.*, a target person (who) performing a distinctive action (what) in a particular location (where). For instance, Figure 1(a) illustrates a 3W INS example where Ian (who) holds a phone (what) in cafe1 (where). These elements form fundamental storytelling units [32], making PAL INS not only contributes to comprehensive video understanding [26], but also an enabling technology for downstream applications such as video question answering [41].

Current 3W INS approaches can be broadly categorized into, *i.e.*, Complete-Decomposition (CD), Non-Decomposition (ND), as

*Both authors contributed equally to this research.
†Corresponding author.

ACM MM, Dublin, Ireland

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX



(a) 3W INS example in *Eastender*

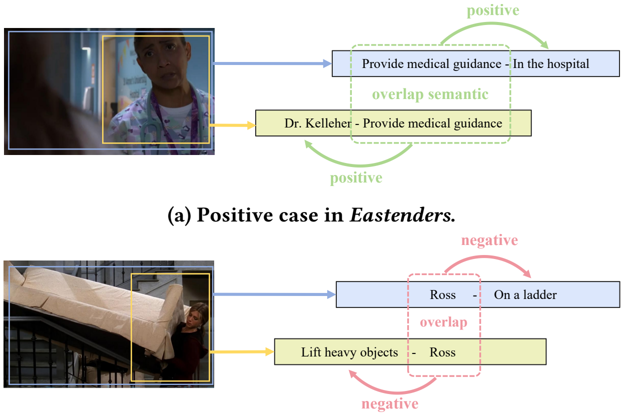
(b) CD method (c) ND method (d) PD method (e) QIPD

Figure 1: Different methods for 3W INS task.

well as Partial-Decomposition (PD) methods. The CD methods (Figure 1(b)) [20, 27, 31] assume semantic independence, decomposing the 3W INS task into three isolated single-semantic retrieval tasks, *i.e.*, Who, What and Where. While efficient, they ignore crucial inter-semantic correlations, limiting retrieval accuracy. The ND methods (Figure 1(c)) [21, 24, 38] treat the 3W composite-semantics as an indivisible whole, leveraging Vision Language Models (VLMs) for end-to-end video-text matching. However, they often fail to capture fine-grained semantic distinctions, particularly for rare or composite concepts. PD methods (Figure 1(d)) [16, 17, 34] strike a balance by decomposing 3W INS into three pairwise 2W INS tasks, *i.e.*, Who-What, What-Where and Where-Who, and aggregate results via product fusion.

Yet, they overlook a critical flaw: semantic overlap among 2Ws. For instance, Who-What and What-Where share the What (action) component. From a probability theory perspective, since these components are not independent, simply multiplying them would yield suboptimal results [1]. The semantic overlap issue can affect the final semantics in multiple ways, with both positive and negative influences. As shown in Figure 2(a), when fusing the two 2W INS results "Dr. kelleher-provide medical advice" and "provide medical advice-in the hospital", the overlapping action semantics of "provide medical advice" enhances the credibility of both "Dr. kelleher" and "in the hospital", since this action typically occurs in hospitals and is performed by doctors. Conversely, Figure 2(b) demonstrates a negative case: of mixing "lift heavy objects-Ross" and "Ross-on a ladder". While these two 2W INS results are individually plausible, fusing them via the overlapping person semantics "Ross" introduces a credibility conflict between the strenuous labor (lift heavy objects) and hazardous setting (on a ladder), ultimately undermining reliability. Therefore, simple multiplicative fusion fails to an accurate 3W INS result when combining three 2WS into 3W. It is necessary to develop a fusion method that can model these diverse influences of semantic overlap.

Quantum interference [13] provides a natural mechanism for diverse multiple influences of overlapping components. When wavefunctions superimpose, their relative phases determine the interference pattern. In particularly, in-phase superposition leads to constructive interference (amplitude enhancement), while anti-phase superposition causes destructive interference (amplitude suppression). Motivated by this mechanism, we employ quantum interference to address the semantic overlap when fusing 2W INS results (Figure 1(d)), where overlapping semantic reinforce or weaken each other through interference. Additionally, considering the varying impacts of shots with different time differences from the current shot, we address this by assigning these shots different phases.


positive
Provide medical guidance - In the hospital
overlap semantic
Dr. Kelleher - Provide medical guidance
positive
negative
Ross - On a ladder
overlap
Lift heavy objects - Ross
negative

(a) Positive case in *Eastenders*.

(b) Negative case in *Friends*.

Figure 2: Positive and negative influence about semantic overlap in story videos.

Specifically, we propose a Quantum Interference Partial Decomposition (QIPD) method for 3W INS. Initially, three types of 2W INS results (Who-What, What-Where, and Where-Who) are extracted from the story videos using the PD method [17]. These results are then processed in the semantic interference module, where they are transformed into amplitude and phase representations and fused through quantum interference to compute the 3W amplitude. Next, the temporal interference module modulates the influence of neighboring shots by assigning large phase differences to distant shots (reducing their impact) and small phase differences to proximate shots (enhancing their correlation). Finally, the final 3W INS results are obtained by combining the amplitude and phase representations of all shots through quantum interference, effectively modeling both semantic correlation and temporal proximity.

In summary, the contributions of this paper include:

- We discover the semantic overlap between the 2Ws, analyze its potential diverse influences, and address this issue using quantum interference theory.
- We propose a comprehensive Quantum Interference Partial Decomposition network featuring: (1) a semantic interference module that dynamically handles overlapping semantics through quantum interference, and (2) a temporal interference module that adaptively weights shot relevance based on temporal proximity.
- Extensive evaluations on 3 public datasets demonstrate the effectiveness and superiority of the proposed QIPD compared to competitive baseline methods.

2 RELATED WORK

2.1 Composite-Semantics Video INS

Composite-semantics video INS task aims to recognize video shots containing specific composite-semantics, enabling structured semantic retrieval of composite-semantics in videos [17]. Existing approaches for composite-semantics video INS can be broadly categorized into three paradigms: CD, ND and PD methods. CD methods [20, 27, 31] assume that composite semantics in the query are mutually independent, thus divide a complex composite-semantics INS problem into multiple single-semantics INS problems [19, 42], ignoring necessary semantic correlations in story videos. ND methods [21, 24, 38] treat composite semantics as an indivisible unit, a paradigm commonly adopted in natural language processing [2, 44]. However, such approaches often face challenges in precise action recognition due to their reliance on generic visual feature representations. PD method [17, 34] offers a strategic advancement by explicitly modeling the correlations among composite semantics in a pairwise division, but they often yield suboptimal results due to overlooked semantic overlaps between the decomposed components. Our proposed QIPD method addresses this issue by modeling semantic overlaps using quantum interference theory.

2.2 Quantum Interference in Multimedia

Quantum theories encompassing quantum interference theory has recently achieved remarkable progress in multimodal domains [15, 25, 40]. Quantum interference, through constructive interference and destructive interference, effectively models both the positive and negative effects within overlapping semantic components. Due to this unique capability, quantum interference has demonstrated significant advantages in fields like humor detection [36] and sentiment classification [35, 46], while also proving effective in sarcasm anaysis [29, 37]. Despite its demonstrated success in these domains, quantum interference remains underexplored for 3W INS tasks.

3 METHOD

Our framework begins by establishing the theoretical foundation of quantum interference. We then formally define the 3W-INS problem, introducing key notations for semantic components and their interactions. The system first processes raw video inputs to extract three distinct 2W instance search results, each producing semantic amplitude outputs. These amplitudes subsequently undergo semantic interference modeling. Finally, temporal interference modulation is applied, generating the optimized 3W-INS results.

3.1 Preliminary

In quantum mechanics, wave ψ is completely described by a complex value:

$$\psi = \rho e^{i\theta} \tag{1}$$

where ρ are real amplitudes and θ are phases.

The probability of the wave follows the Born rule, which states that the probability P equals the squared of its amplitude:

$$P = \rho^2 \tag{2}$$

Quantum interference stems from the superposition principle in quantum mechanics. When two coherent waves $\psi_1 = \rho_1 e^{i\theta_1}$ and $\psi_2 = \rho_2 e^{i\theta_2}$ interact, their amplitudes superimpose. The resultant amplitude ρ_{12} of is given by:

$$\begin{aligned}\rho_{12} &= |\psi_1 + \psi_2| \\ &= \left|\rho_1 e^{i\theta_1} + \rho_2 e^{i\theta_2}\right| \\ &= \sqrt{\rho_1^2 + \rho_2^2 + 2\rho_1\rho_2\cos(\Delta\theta_{1,2})} \\ &\equiv \sqrt{\rho_1^2 + \rho_2^2 + i_{1,2}} \quad (i_{j,k} = 2\rho_j\rho_k\cos\Delta\theta_{i,j})\end{aligned} \tag{3}$$

The interference term $i_{1,2}$ explicitly depends on the phase difference $\Delta\theta_{1,2} = \theta_1 - \theta_2$ as well as the amplitudes ρ_1 and ρ_2.

Thus, due to quantum interference effects, the probability p_{12} resulting from the superposition of wave ψ_1 and wave ψ_2 is given by:

$$P_{12} = P_1 + P_2 + i_{1,2} \tag{4}$$

where P_1 and P_2 denote the probability of wave ψ_1 and wave ψ_2. When the interference term is zero, the quantum probability superposition reduces to the classical linear addition of probability. When the interference term is positive, the resultant probability increases; when the interference term is negative, the resultant probability decreases.

For N-state systems, this extends to pairwise interference:

$$|\sum_{j=1}^{n}\psi_j|^2 = \sum_{j=1}^{n}\rho_j{}^2 + \sum_{1\le j<k\le n} i_{j,k} \tag{5}$$

3.2 Formulation

For the 3W INS task, the goal is to find video shots of p-th person at l-th location performing a-th action . To accomplish this task, we need to obtain the score for each video shot regarding this instance.

In the 2W INS module, we use the PD method [17] to compute three 2W INS score matrix: person-action $\mathbf{S}^{PA} \in \mathbb{R}^{|P|\times|A|}$, action-location $\mathbf{S}^{AL} \in \mathbb{R}^{|A|\times|L|}$, and location-person $\mathbf{S}^{LP} \in \mathbb{R}^{|L|\times|P|}$, where $|P|$, $|A|$ and $|L|$ denote the total number of the person, action and location categories in the query sets. The score matrices contain instance-specific probabilities, where $P_{pa} \in \mathbb{R}$, $P_{al} \in \mathbb{R}$, $P_{lp} \in \mathbb{R}$ are the 2W INS scores of p-th person and a-th action , a-th action and l-th loction as well as l-th location and p-th person. In the semantic interference module, these 2W probabilities are used to derive the 2W semantic phases $\theta_{pa} \in \mathbb{R}$, $\theta_{al} \in \mathbb{R}$ and $\theta_{lp} \in \mathbb{R}$ and amplitudes $\rho_{pa} \in \mathbb{R}$, $\rho_{al} \in \mathbb{R}$ and $\rho_{lp} \in \mathbb{R}$, which are then used to construct the 3W semantics amplitude $\rho_{pal} \in \mathbb{R}$. In the temporal interference module, We need to consider other shots, the number of shots is N, and the current shot is n-th shot. we realign the amplitudes $\{\rho'_m\}_{m=1}^{N} \in \mathbb{R}^N$ and phases $\{\theta'_m\}_{m=1}^{N} \in \mathbb{R}^N$ of nearby shots and ultimately output the 3W semantic amplitude $\bar{\rho}'_{n,pal}$ fused with other shots. The square of this amplitude $P_{n,pal}$ serves as the final score for the n-th video shot.

3.3 2W INS

We follow [17], using Person-Action (PA) module $\mathbf{F}^{PA}$, Action-Location (AL) module $\mathbf{F}^{AL}$, and Location-Person (LP) module $\mathbf{F}^{LP}$ to obtain different 2W INS score matrices:

$$\mathbf{S}^{PA} = \mathbf{F}^{PA}(v), \quad \mathbf{S}^{AL} = \mathbf{F}^{AL}(v), \quad \mathbf{S}^{LP} = \mathbf{F}^{LP}(v). \tag{6}$$

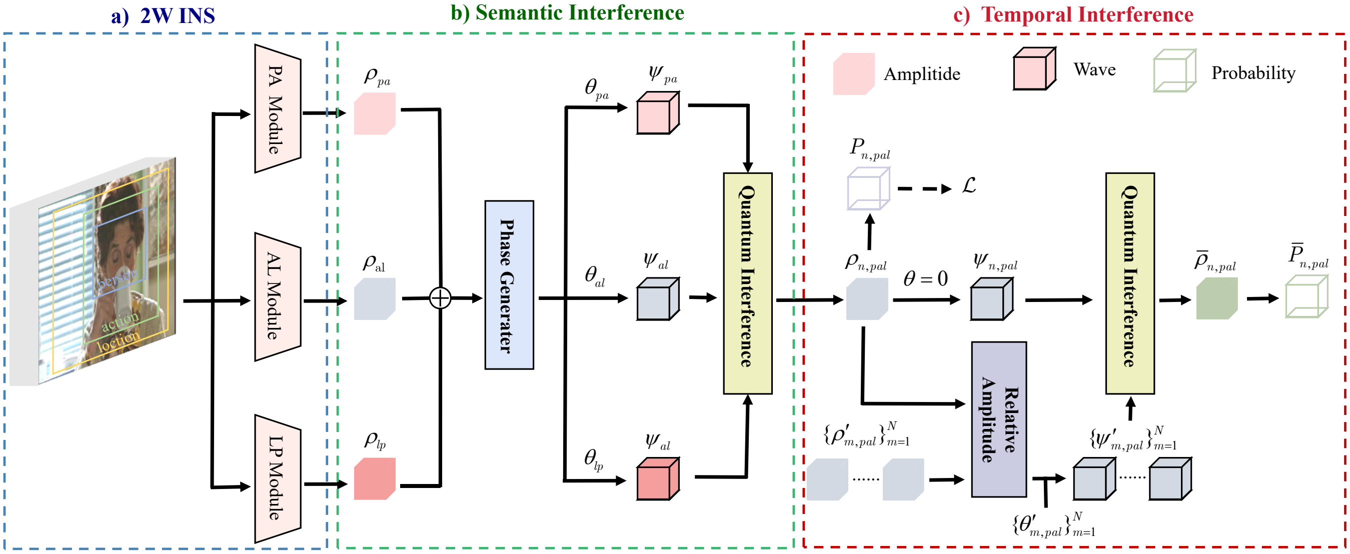


Figure 3: Model architecture. We model semantic overlap using the semantic interference module and distinguish the effects of different shots with the temporal interference module.

where v denotes the current video shot and $\mathbf{S}^{PA}$, $\mathbf{S}^{AL}$ and $\mathbf{S}^{AL}$ denote the score matrices of three 2W INS results.

These score matrices have been normalized, where the different scores in each matrix can be interpreted as probabilities for distinct instances:

$$\mathbf{S}^{PA} = \{P_{pa}\}_{p=1,a=1}^{|P|,|A|},\ \mathbf{S}^{AL} = \{P_{al}\}_{a=1,l=1}^{|A|,|L|},\ \mathbf{S}^{LP} = \{P_{lp}\}_{l=1,p=1}^{|L|,|P|}. \tag{7}$$

where P_{pa}, P_{al} and P_{lp} represent the probability of the p-th person and a-th action, the a-th person and l-th action as well as the l-th person and p-th action.

3.4 Semantic Interference

After obtaining the three 2W INS results, we need to fuse them to derive the 3W INS result. We employ quantum interference (QI) module to fuse the three 2W INS results. According to the relationship between probability and amplitude, the probability can be calculated as:

$$\rho_{pa} = \sqrt{P_{pa}}, \quad \rho_{al} = \sqrt{P_{al}}, \quad \rho_{lp} = \sqrt{P_{lp}}. \tag{8}$$

where ρ_{pa}, ρ_{al} and ρ_{lp} denote the amplitudes between person p and action a, action a and location l, as well as location l and person p, respectively.

At this stage, constructing the wave still requires the corresponding phase. To address this, we design a phase generator. We concatenate these amplitudes and pass them through a Multilayer Perceptron (MLP) to obtain the phase.

$$\theta_{pa}, \theta_{al}, \theta_{lp} = \pi \cdot \tanh\left(\mathrm{MLP}\big(\mathrm{concate}(\rho_{pa}, \rho_{al}, \rho_{lp})\big)\right) \tag{9}$$

where θ_{pa}, θ_{al} and θ_{lp} denote the phrases, concate$(\cdot)$ denotes the concatenate operation, and MLP denote the Multilayer Perceptron. To ensure the phase adheres to physical constraints, we map it to the interval $(-\pi, \pi)$ using $\pi \cdot \tanh(\cdot)$.

With all magnitude and phase components obtained, we can obtain the waves corresponding to each 2W INS result

$$\psi_{pa} = \rho_{pa} e^{i\theta_{pa}}, \quad \psi_{al} = \rho_{al} e^{i\theta_{al}}, \quad \psi_{lp} = \rho_{lp} e^{i\theta_{lp}}. \tag{10}$$

We model the final 3W wave as a superposition of 2W waves, using their interference terms to capture semantic overlap. Thus, the final amplitude ρ_{pal} can be computed by leveraging quantum interference theory, formulated as:

$$\rho_{pal} = |\psi_{pa} + \psi_{al} + \psi_{lp}| \tag{11}$$

We use the probability of the interfered wave, *i.e.*, the squared magnitude of the amplitude, as the final score for ternary semantic relationships:

$$P_{pal} = {\rho_{pal}}^2 \tag{12}$$

The final loss $\mathcal{L}$ is computed using the Mean Squared Error (MSE) loss, formulated as:

$$\mathcal{L} = \mathrm{MSE}(gt_{pal}, P_{pal}) \tag{13}$$

3.5 Temporal Interference

After completing the model training, we employ quantum interference theory to perform rank optimization across different shots. Since other shots are involved, we assume there are N shots , with the current one being the n-th shot, and the results from different shots are distinguished using subscripts.

We estimate the semantic affinity degree $w_{n,m}$ between the n-th shot and m-th shot, which is calculated by:

$$w_{n,m} = \theta \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{|n-m|^2}{2\sigma^2}\right) \cdot e_{n,m} \tag{14}$$

where θ is a coefficient to modulate the semantic affinity and σ is the standard deviation of a standard normal distribution. And $e_{n,m}$ represents the visual similarity between the n-th video shot and the m-th video shot.

We first compute the relative amplitudes of other shots. The amplitude differences, weighted by their semantic affinity degrees, are then adopted as the new relative amplitudes. Since we only consider the positive influence of other shots on the amplitude, any relative amplitude with a negative amplitude difference is set to 0:

$$\rho'_{m,pal} = w_{n,m} \cdot \max(\rho_{m,pal} - \rho_{n,pal}, 0) \tag{15}$$

where $\max(\cdot,\cdot)$ denotes maximization operation.

To differentiate the influence of adjacent shots versus distant shots, we assign phase terms based on their temporal distance from the current shot, where the current shot serves as the reference with zero phase while other shots are assigned a phase proportional to their shot index difference multiplied by a phase coefficient k:

$$\psi'_{m,pal} = \rho'_{m,pal} \cdot e^{k(n-m)i} \tag{16}$$

where $\psi_{m,pal}$ denotes the m-th shot's wave, and k denote phase coefficient with a value of 0.01 which ensures phase differences remain bounded within $(-\pi, \pi)$.

The phase of the n-th shot is set to 0, and its wave function $\psi_{n,pal}$ is directly represented by its amplitude $\rho_{n,pal}$. The final assignment of 3W INS semantics $\bar{\rho}_{n,pal}$ is obtained through quantum interference-based superposition of wave functions from all shots:

$$\bar{\rho}_{n,pal} = |\psi_{n,pal} + \sum_{m=1}^{N} \psi'_{m,pal}| \tag{17}$$

The final 3W probability $\bar{P}_{pal}$ is also calculated by converting the amplitudes to probabilities through quantum theory, where probabilities equal the square of the amplitudes. The probability then serves as the final 3W INS score for shot ranking:

$$\bar{P}_{pal} = \bar{\rho}^2_{n,pal} \tag{18}$$

4 EXPERIMENT

To validate the superiority of the proposed QIPD method, we conduct experimental evaluations on three large-scale 3W INS datasets. We compare our proposed method QIPD with both SOTA 3W INS approaches and Rank Aggregation (RA) methods Then, the effectiveness of individual parts is analyzed in QIPD method through the ablation study, and the dynamic performance of QIPD method is investigated with varying model parameters. Lastly, qualitative results of different cases in 3W INS datasets demonstrate the real performance of various 3W INS methods.

4.1 Experiment Settings

Datasets. We use three 3W INS datasets [17] built from two main resources: (1) one British television soap opera Eastenders, (2) two famous American TV series including Friends and The Big Bang Theory (TBBT) [17]. Three large-scale 3W INS datasets totally comprise 86 search topics spanning 675,759 shots from three TV series, including Eastenders, Friends and TBBT. Specifically, the Eastenders consists of 43 topics including 12 persons, 6 actions and 8 locations, with 1,567 groundtruths. The Friends comprises 43 topics with 463 groundtruths, composed of 8 persons, 4 actions and 7 locations. The TBBT contains 22 topics involving 9 persons, 5 actions and 6 locations, with 451 groundtruths. Generally, the number of groundtruth shots in the training set is higher than that in the test set, and the overlap of semantics is minimized to prevent any 3W INS topics from appearing in both sets.

Evaluation Metrics Similar to [34], we adopt the mean average precision (mAP) to evaluate the overall performance across all queried topics:

$$\text{mAP} = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{i \in \mathcal{R}_q} P@i(q)}{|\mathcal{R}_q|} \tag{19}$$

where Q is the entire query set, $\mathcal{R}_q$ is the set of positions of the answer shots for query topic q in the corresponding ranked list, and $P@i(q)$ is the precision of the top-i shots in the ranked list corresponding to query topic q.

Implementation Details. We use Retinaface [9] pre-trained on WIDER Face dataset [9] to detect faces in videos, and Arcface [8] pre-trained on MS1Mv2 dataset [18] to recognize detected faces. PPDM [28], pre-trained on HICO-DET dataset [5], is used for action detection and recognition. SASR [30] is used to extract visual features for location recognition. To train the QIPD network, we adopt the Adam optimizer with a learning rate of 0.0001 and a dropout rate of 0.5. We set model paremeter $\theta = 1.6$ and $\sigma = 7$.

4.2 Comparison with 3W INS methods.

Comparative Baselines. To evaluate the effectiveness of the proposed QIPD, we select the following related works for quantitative comparation:

- **P×A×L** [17] denote the typical CD method, which multiplies the recognition scores of independent semantics.
- **ALBEF** [24] introduces "momentum distillation", leading to better performance by creating more coherent and synergistic representations of the two data types.
- **CLIP** [38] explores training visual models using BERT [10] as supervision, leveraging the semantic information in language to guide the learning process of ViT [11].
- **ALPRO** [21] introduces "entity prompt" to align video content with relevant language entities. It significantly improves performance in tasks requiring the interpretation and association of video content with language.
- **BLIP and BLIP-2** [22, 23] employ a method that bootstraps the model with rich, multimodal information. The BLIP-2 innovates by combining frozen image encoders with Large-scale Language Models (LLM).
- **Vitamin** [6] explores the effectiveness of different structures in VLM and proposes a 3-stage hybrid architecture.
- **SAM** [45] uses the visual features of actions and locations using an attention scheme to generate a concerted feature for composite-semantics retrieval.
- **3W** [17] partially decompose the 3W INS problem into three semantic-correlated 2W INS problems *i.e.*, person-action INS, action-location INS, and location-person INS.

Results on Eastenders. The 3W INS results on Eastemders dateset are shown in Table 1, where the red number represents the optimal value in each column, while the blue number indicates the second-best value. This notation is consistently applied in all subsequent tables. In the Eastenders dataset, We can see that QIPD method exhibits a remarkable mAP performance of 22.8%, which is higher than both CD method and the best ND method ViTamin.

Table 1: The comparative results (%) of 3W INS methods on Eastenders.

Method		Venue	Query Topic ID																					mAP
			13	16	19	20	26	28	30	36	37	39	40	44	46	48	49	50	51	52	55	56	57	
CD	P× A× L [17]	TIP 25	42.4	20.1	18.6	31.9	1.7	8.0	20.4	1.0	13.6	13.7	14.3	17.8	19.7	11.0	20.2	0.4	1.4	4.3	55.0	0.7	18.1	15.9
ND	ALBEF [24]	NIPS 21	17.2	5.8	3.6	6.4	1.1	2.1	11.5	0.2	6.3	1.5	10.6	11.0	4.6	9.5	5.1	0.1	0.4	1.7	8.1	0.4	3.3	5.3
	CLIP [38]	ICML 21	21.2	6.6	4.1	7.7	1.0	2.1	12.1	0.2	6.6	1.7	6.4	20.2	3.9	8.2	5.7	0.1	0.5	1.9	9.2	0.4	3.3	5.9
	ALPRO [21]	CVPR 22	30.5	7.6	5.9	9.4	1.0	2.8	11.1	0.3	6.4	2.2	6.1	17.6	5.1	9.9	7.4	0.1	0.5	2.2	18.4	0.4	3.6	7.1
	BLIP [23]	ICML 22	31.6	7.6	5.0	10.1	1.0	2.8	10.1	0.4	6.3	2.0	4.5	10.4	5.2	10.8	8.9	0.2	0.7	2.4	30.4	0.4	3.6	7.3
	BLIP-2 [22]	ICML 23	30.1	7.7	5.0	10.6	1.1	2.9	9.0	0.5	6.6	1.9	8.1	9.7	4.1	20.1	7.3	0.2	0.6	2.2	20.6	0.4	3.9	7.3
	ViTamin [6]	CVPR 24	18.8	17.3	5.9	12.3	1.9	6.2	4.3	9.2	2.4	10.4	3.7	1.3	4.1	3.1	3.5	2.7	4.2	0.8	19.4	0.2	23.7	7.4
PD	SAM [45]	TMM 21	42.1	20.1	18.4	34.8	3.1	7.6	6.6	1.1	15.0	4.0	27.0	12.0	12.5	11.2	20.2	2.1	1.4	3.6	58.2	0.7	19.4	15.3
	3W [17]	TIP 25	49.8	33.1	15.9	45.9	2.1	7.8	30.4	1.1	20.6	21.4	22.5	15.6	24.0	23.1	47.8	0.1	1.2	3.1	71.9	0.9	18.7	21.6
	QIPD	Ours	**51.2**	**36.4**	**18.7**	**46.0**	**2.5**	**10.9**	**20.8**	**0.8**	**13.8**	**25.4**	**22.5**	**27.3**	**14.4**	**27.7**	**46.3**	**0.3**	**2.0**	**4.8**	**80.7**	**0.3**	**21.5**	**22.8**

Table 2: The comparative results (%) of 3W INS methods on Friends and TBBT.

Method		Venue	Friends Query Topic ID												mAP	TBBT Query Topic ID											mAP
			11	14	32	37	46	47	49	51	52	59	60	61		36	37	39	42	45	50	52	53	54	58	62	
CD	P×A×L [17]	TIP 25	3.7	2.7	22.1	0.4	0.5	0.6	2.3	11.9	6.8	0.8	0.7	10.6	5.3	0.0	0.0	1.9	0.9	2.3	4.3	9.0	1.1	2.6	5.5	2.0	2.7
ND	ALBEF [24]	NIPS 21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
	CLIP [38]	ICML 21	0.5	0.0	1.6	0.0	0.0	0.0	0.0	0.6	0.2	0.0	0.0	6.4	0.8	0.1	0.0	0.5	0.1	0.9	0.0	0.0	1.2	1.4	10.8	0.0	1.4
	ALPRO [21]	CVPR 22	0.3	0.0	0.6	0.0	0.0	0.0	0.1	0.3	0.1	0.0	0.0	5.6	0.6	0.0	0.0	11.3	0.0	0.1	0.0	0.0	0.1	0.4	5.6	1.4	1.7
	BLIP [23]	ICML 22	1.4	0.1	7.6	0.0	0.0	0.0	0.2	2.4	0.5	0.1	0.2	9.6	1.8	0.1	0.0	9.8	4.5	3.1	0.5	0.2	1.2	2.6	7.0	0.7	2.7
	BLIP-2 [22]	ICML 23	1.1	0.1	7.0	0.1	0.0	0.1	0.2	1.8	0.8	0.0	0.1	9.0	1.7	0.1	0.0	8.8	2.0	3.2	0.6	0.2	0.9	2.3	6.5	0.0	2.2
	ViTamin [6]	CVPR 24	0.6	0.1	7.9	0.0	0.0	0.0	0.1	1.3	0.4	0.1	0.0	7.4	1.5	0.0	0.0	3.9	2.9	2.3	7.7	4.0	0.9	3.6	4.9	4.4	3.1
PD	SAM [45]	TMM 21	2.7	1.1	0.1	0.1	0.2	0.1	0.1	4.0	1.5	0.1	0.0	7.7	1.5	0.0	0.0	0.0	0.4	2.3	1.8	1.9	0.1	0.4	1.0	0.0	0.7
	3W [17]	TIP 25	4.9	4.3	55.8	0.7	0.7	0.9	7.0	17.8	6.1	1.1	0.3	10.2	9.1	0.0	0.0	7.9	3.1	2.3	8.9	3.8	0.7	7.2	4.5	8.7	4.3
	QIPD	Ours	**4.7**	**3.4**	**62.4**	**0.6**	**0.7**	**0.8**	**7.3**	**17.3**	**9.7**	**1.0**	**1.2**	**13.7**	**10.2**	**0.0**	**0.0**	**21.1**	**3.5**	**2.3**	**8.9**	**4.3**	**1.2**	**11.2**	**12.8**	**12.4**	**7.1**

Results on Friends and TBBT. The 3W INS results on Friends and TBBT datesets are shown in Table 2. On the Friends dataset, QIPD achieves the highest mAP score of 10.2%, surpassing the 4.9% of the conventional CD method P×A, and significantly exceeding the 8.5% of the best ND method BLIP. On The Big Bang Theory (TBBT) dataset, QIPD also demonstrates superior performance, achieving an mAP score of 7.1%. This result nearly doubles the 4.4% mAP of the conventional CD method P×A and significantly outperforms the 4% mAP attained by the best ND method, ViTamin. Compared with the PD method, our proposed QIPD achieves a 1.2% improvement, effectively validating the efficacy of quantum interference modeling. Compared with the PD method, QIPD achieves improvements of 1.1% and 2.8% on these two datasets, respectively.

Additionally, we observe a phenomenon: QIPD achieves significantly greater performance improvement on the TBBT dataset compared to the other two datasets. In TBBT, the main characters (e.g., Sheldon) exhibit distinctive professional traits (as physicists), which creates strong correlations between their identities, settings (e.g., laboratories), and actions (e.g., scientific activities). This results in more pronounced semantic overlap within the data. Consequently, our method demonstrates more substantial performance gains.

4.3 Comparison with RA methods.

RA methods. For systematic validation of our quantum interference fusion method, we establish baseline rankings using three 2W INS results generated by 3w and conduct comparative experiments with different RA methods, which contains: RRF [7], PostNDCG [14], Median [12], MEAN [4], HPA [14], Dowdall [39] , ER [33], CG [43] and BordaCounT [3].

Results on Eastenders. Table 3 presents the comparison of 3W INS results between QIPD and different RA methods on the Easter dataset. 3W w/ A represents the baseline rank generated by 3W, followed by rank aggregation using Method A. QIPD outperforms the best ranking fusion method HPA by 1.3%, demonstrating the superiority of our approach. It is worth noting that these RA methods fail to surpass PD, a simple product-based fusion method, indicating that fusion approaches disregarding the characteristics of 3W INS cannot achieve satisfactory results.

Results on Friends and TBBT. Table 4 compares the 3W INS results of QIPD and different RA methods on the Friends and TBBT datasets. QIPD outperforms the best ranking fusion method, MEAN, by 0.5% and 2.9%, further demonstrating its superiority. Notably, the top-performing RA method, MEAN, simply averages predictions, suggesting that complex approaches do not necessarily surpass well-designed simple methods, *e.g.*, MEAN and our proposed QIPD in 3W INS tasks.

4.4 Ablation Study

Baselines. To systematically evaluate the contributions of semantic interference (SI) and temporal interference (TI) in our QIPD framework, we conduct an ablation study comparing four variants on 3W INS datasets: (1) The baseline 3W method without either SI or TI modules; (2) 3W w/ SI incorporating only semantic interference; (3) 3W w/ TI with solely temporal interference; and (4) our complete QIPD framework integrating both CI and TI modules

Table 3: The comparative results (%) of RA methods on Eastenders.

Method	Query Topic ID																					mAP
	13	16	19	20	26	28	30	36	37	39	40	44	46	48	49	50	51	52	55	56	57	
3W	49.8	33.1	15.9	45.9	2.1	7.8	30.4	1.1	20.6	21.4	22.5	15.6	24.0	23.1	47.8	0.1	1.2	3.1	71.9	0.9	18.7	21.6
3W w/ RRF	0.0	0.0	0.0	0.0	16.7	2.6	0.1	0.0	0.1	0.3	0.0	0.0	0.1	0.0	0.0	0.8	0.1	0.2	0.0	5.9	0.0	1.3
3W w/ PostNDCG	0.0	0.0	0.2	0.0	0.1	0.6	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.5	0.0	0.0	0.1	0.1
3W w/ Median	0.0	0.0	0.0	0.0	16.7	2.6	0.1	0.0	0.1	0.3	0.0	0.0	0.1	0.0	0.0	0.8	0.1	0.2	0.0	5.9	0.0	1.3
3W w/ MEAN	39.5	46.2	23.9	20.1	2.7	10.6	13.6	1.4	7.2	28.6	25.3	4.9	30.5	20.2	47.7	0.8	1.5	2.9	81.0	0.3	31.9	21.0
3W w/ HPA	42.6	46.3	24.0	24.7	2.7	10.1	15.5	1.5	7.2	28.7	28.4	5.7	31.4	20.9	45.8	0.8	1.7	4.1	75.9	0.4	32.4	21.5
3W w/ Dowdall	38.8	46.0	23.8	19.1	2.7	10.4	13.1	1.4	7.0	28.7	25.2	4.7	30.6	19.9	47.6	0.6	1.4	2.8	81.0	0.3	31.8	20.8
3W w/ ER	39.5	46.2	23.7	20.0	2.7	10.6	13.6	1.5	7.2	28.6	25.3	4.9	30.5	20.2	47.7	0.8	1.5	2.9	82.1	0.3	31.9	21.0
3W w/ CG	39.5	46.2	23.9	20.1	2.7	10.6	13.6	1.4	7.2	28.6	25.3	4.9	30.5	20.2	47.7	0.8	1.5	2.9	81.0	0.3	31.9	21.0
3W w/ BordaCount	0.0	0.0	0.0	0.0	16.7	2.6	0.1	0.0	0.1	0.3	0.0	0.0	0.1	0.0	0.0	0.8	0.1	0.2	0.0	5.9	0.0	1.3
QIPD	**51.2**	**36.4**	**18.7**	**46.0**	**2.4**	**10.9**	**20.8**	**0.8**	**13.8**	**25.4**	**27.3**	**14.4**	**27.7**	**26.7**	**46.3**	**0.3**	**2.0**	**4.8**	**80.7**	**0.3**	**21.5**	**22.8**

Table 4: The comparative results (%) of RA methods on Friends and TBBT.

Method	Friends Query Topic ID												mAP	TBBT Query Topic ID											mAP
	11	14	32	37	46	47	49	51	52	59	60	61		36	37	39	42	45	50	52	53	54	58	62	
3W	4.9	4.3	55.8	0.7	0.7	0.9	7.0	17.8	6.1	1.1	0.3	10.2	9.1	0.0	0.0	7.9	3.1	2.3	8.9	3.8	0.7	7.2	4.5	8.7	4.3
3W w/ RRF	1.2	1.5	0.0	0.0	0.0	0.3	0.1	0.0	0.0	0.0	0.0	9.7	1.1	0.0	0.0	0.0	0.0	2.3	0.0	1.2	0.0	0.2	4.4	1.3	0.9
3W w/ PostNDCG	0.1	0.0	0.1	0.0	0.0	0.1	0.0	0.1	0.3	0.0	0.1	4.7	0.5	0.0	0.0	0.0	0.0	2.3	0.2	1.2	0.0	0.1	1.8	0.4	0.5
3W w/ Median	1.2	1.5	0.0	0.0	0.0	0.3	0.1	0.0	0.0	0.0	0.0	9.7	1.1	0.0	0.0	0.0	0.0	2.3	0.0	1.2	0.0	0.2	4.4	1.3	0.9
3W w/ MEAN	7.3	3.9	52.9	0.8	2.8	0.9	2.1	23.3	8.6	1.1	1.3	11.4	9.7	0.0	0.0	9.4	3.3	2.3	9.4	3.8	1.0	6.1	4.9	5.7	4.2
3W w/ HPA	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	8.2	0.7	0.0	0.0	0.0	0.1	0.6	0.2	1.2	0.0	0.0	4.9	0.2	0.7
3W w/ Dowdall	7.3	3.9	52.8	0.7	2.8	0.9	2.1	23.2	8.6	1.1	1.3	11.7	9.7	0.0	0.0	9.3	3.2	2.3	9.3	3.7	0.9	6.0	4.6	5.5	4.1
3W w/ ER	7.3	3.9	52.9	0.7	2.8	0.9	2.1	23.3	8.6	1.1	1.3	11.3	9.7	0.0	0.0	9.4	3.3	2.3	9.4	3.8	1.0	6.1	4.9	5.7	4.2
3W w/ CG	7.3	3.9	52.9	0.8	2.8	0.9	2.1	23.3	8.6	1.1	1.3	11.4	9.7	0.0	0.0	9.4	3.3	2.3	9.4	3.8	1.0	6.1	4.9	5.7	4.2
3W w/ BordaCount	1.2	1.5	0.0	0.0	0.0	0.3	0.1	0.0	0.0	0.0	0.0	9.7	1.1	0.0	0.0	0.0	0.0	2.3	0.0	1.2	0.0	0.2	4.4	1.3	0.9
QIPD	**4.7**	**3.4**	**62.4**	**0.6**	**0.7**	**0.8**	**7.3**	**17.3**	**9.7**	**1.0**	**1.2**	**13.7**	**10.2**	**0.0**	**0.0**	**21.1**	**3.5**	**2.3**	**8.9**	**4.3**	**1.2**	**11.2**	**12.8**	**12.4**	**7.1**

Table 5: Ablation Study Results on Three Datasets

Method	Eastenders	Friends	TBBT	Average
3W	21.6	9.1	4.3	11.7
3W w/ CI	22.3	9.8	5.2	12.4
3W w/ TI	22.7	10.1	6.9	13.2
QIPD	**22.8**	**10.2**	**7.1**	**13.4**

for joint quantum interference modeling. This controlled ablation study enables precise measurement of each component's individual and combined effects on overall system performance.

Results. As presented in Table 5, our ablation study reveals significant performance gains through quantum interference modeling. While the baseline 3W achieves 11.7% mAP, incorporating content interference (3W w/ SI) and temporal interference (3W w/ TI) improves performance by 0.7% and 1.5% respectively, demonstrating the individual value of both modalities. The complete QIPD framework, integrating both SI and TI, achieves the highest mAP of 13.4%, validating the synergistic effect of quantum interference.

4.5 Dynamic Performance

We conducted dynamic performance experiments on parameters θ and σ using the Eastender dataset. Figure 5 presents the results of the ablation experiment with respect to the coefficient to modulate the semantic affinity θ and the standard deviation of a standard normal distribution σ. Our experimental results demonstrate that the proposed QIPD method achieves robust performance across a wide range of parameter combinations, with the optimal accuracy of 22.8% attained at $\theta = 1.6$ and $\sigma = 7$. As shown in Figure 5, all parameter values fall within a narrow range of 21.9% to 22.8%, with a minimal variation span of only 0.9%. 83.3% of the parameters are concentrated between 22.6% and 22.8%, demonstrating the method's robust stability across parameter choices. This tight distribution suggests that the method performs consistently well without requiring precise parameter tuning in practice.

4.6 Efficiency Analysis

Table 6 compares model efficiency and performance, where time is inference time per sample, # Para. denotes total trainable parameters and GFLOPS measures giga floating point operations per sample.

Our proposed QIPD achieves the best performance (22.8%) while maintaining competitive efficiency. With 153M parameters and 29.3 GFLOPS, it demonstrates a favorable trade-off between accuracy and computational cost. Compared to the similarly efficient 3W , QIPD improves accuracy by 1.2% with only a marginal increase in inference time. Notably, QIPD significantly outperforms larger models like BLIP (7.3% mAP, 553M params) and ALPRO (7.1% mAP, 337M params) while being 2.2–3.6× more parameter-efficient. The results highlight QIPD's effectiveness in balancing high performance and low computational overhead.

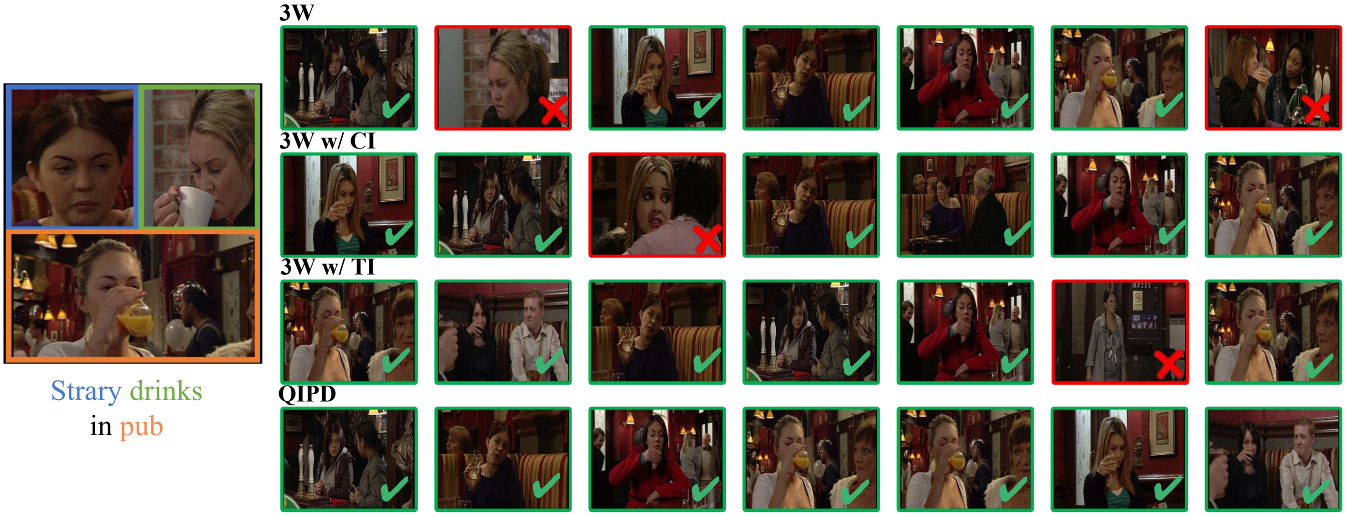


Figure 4: Qualitative comparison resultsin Eastenders. The correct shots are enclosed in green with incorrect ones in red.

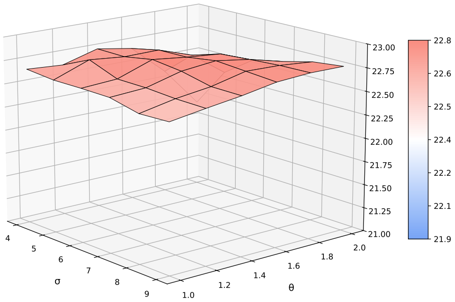


Figure 5: Dynamic Performance of parameters θ and σ.

Table 6: Comparative analysis of model efficiency.

Method	mAP	# Para.	Time	GFLOPS
ALBEF	5.3	316M	15.8s	17.6
CLIP	5.9	256M	3.7s	16.9
ALPRO	7.1	337M	3.3s	96.1
BLIP	7.3	553M	5.6s	191.2
SAM	15.3	128M	1.3s	22.6
3W	21.6	151M	1.9s	29.3
QIPD	22.8	153M	2.0s	29.3

4.7 Qualitative Results

Figure 4 shows a search topic and the corresponding top-7 ranking results generated by 3W, 3W with context interference (3W w/ CI), 3W with temporal interference (3W w/ CI) and our proposed QIPD methods, respectively. The left side displays the queried person, action and location. The right side presents the top-7 shots, where correct shots are enclosed in green and incorrect ones in red.

In the case of "Starry drinks in pub", the 3W yields suboptimal retrieval results due to its inability to model semantic overlap. The 3W w/ CI addresses this linguistic overlap issue and consequently improves retrieval performance. Similarly, 3W w/ TI, which considers the influence of adjacent video segments, also demonstrates performance enhancement. Our proposed QIPD method, which incorporates both types of interference (context and temporal), achieves superior retrieval results.

5 CONCLUSION

In this paper, we proposed Quantum Interference Partial Decomposition (QIPD), a novel approach for 3W INS in story videos, addressing the critical semantic overlap challenge in Partial Decomposition (PD) methods. Inspired by quantum interference theory, our method models semantic reinforcement and conflict through constructive and destructive interference, while temporal interference dynamically weights shot relevance based on proximity. Extensive experiments on three public datasets demonstrate that QIPD outperforms existing methods, validating its effectiveness in fine-grained video understanding.

The current work is still limited to TV datasets. We prioritized TV dramas because they provide rich script descriptions and subtitle timings, making it easier to extract 3W elements. For the same reasons, other video types with abundant textual descriptions, such as movies, stage plays, and sports videos, are also suitable. Such data could be used to evaluate our method if 3W annotations are available, making this a viable direction for future work.

6 Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62372339, 62476021, 72434005, 62371350, 62372336) and the Fundamental Research Funds for the Central Universities (No.2042023kf1033). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Robert B Ash and Catherine A Doléans-Dade. 2000. *Probability and measure theory*. Academic press.

[2] Siamak Barzegar, Andre Freitas, Siegfried Handschuh, and Brian Davis. 2017. Composite semantic relation classification. In *Natural Language Processing and Information Systems: 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Liège, Belgium, June 21-23, 2017, Proceedings 22*. Springer, 406–417.

[3] JC de Borda. 1781. M'emoire sur les' elections au scrutin. *Histoire de l'Acad'emie Royale des Sciences* (1781).

[4] Christopher Burges, Krysta Svore, Paul Bennett, Andrzej Pastusiak, and Qiang Wu. 2011. Learning to rank using an ensemble of lambda-gradient models. In *Proceedings of the learning to rank Challenge*. PMLR, 25–35.

[5] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*. 1017–1025.

[6] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. 2024. Vitamin: Designing scalable vision models in the vision-language era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12954–12966.

[7] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 758–759.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[9] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641* (2019).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[12] Ronald Fagin, Ravi Kumar, and Dandapani Sivakumar. 2003. Efficient similarity search and classification via rank aggregation. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. 301–312.

[13] Richard P. Feynman. 1948. Space-Time Approach to Non-Relativistic Quantum Mechanics. *Reviews of Modern Physics* 20, 2 (1948), 367–387. doi:10.1103/RevModPhys.20.367

[14] Soichiro Fujita, Hayato Kobayashi, and Manabu Okumura. 2020. Unsupervised ensemble of ranking models for news comments using pseudo answers. In *European Conference on Information Retrieval*. Springer, 133–140.

[15] Dimitrios Gkoumas, Qiuchi Li, Yijun Yu, and Dawei Song. 2021. An entanglement-driven fusion neural network for video sentiment analysis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 1736–1742.

[16] Jiahao Guo, Chao Liang, and Zhongyuan Wang. 2023. Who, What and Where: Composite-semantic Instance Search for Story Videos. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE Computer Society, 858–863.

[17] Jiahao Guo, Ankang Lu, Zhengqian Wu, Zhongyuan Wang, and Chao Liang. 2025. Who, What and Where: Composite-Semantics Instance Search for Story Videos. *IEEE Transactions on Image Processing* (2025).

[18] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 87–102.

[19] Martin Höffernig and Werner Bailer. 2016. JOANNEUM RESEARCH at TRECVID 2016 Instance Search Task.. In *TRECVID*.

[20] Hervé Le Borgne. 2017. IRIM at TRECVID 2017: Instance Search. In *TRECVID*.

[21] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2022. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4953–4963.

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[25] Qiuchi Li, Dimitris Gkoumas, Christina Lioma, and Massimo Melucci. 2021. Quantum-inspired multimodal fusion for video sentiment analysis. *Information Fusion* 65 (2021), 58–71.

[26] Ruizhe Li, Jiahao Guo, Mingxi Li, Zhengqian Wu, and Chao Liang. 2023. A Hierarchical Deep Video Understanding Method with Shot-Based Instance Search and Large Language Model. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9425–9429.

[27] Ya Li, Guanyu Chen, Xiangqian Cheng, Chong Chen, Shaoqiang Xu, Xinyu Li, Xuanlu Xiang, Yanyun Zhao, Zhicheng Zhao, and Fei Su. 2019. BUPT-MCPRL at TRECVID 2019: ActEV and INS.. In *TRECVID*.

[28] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. 2020. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 482–490.

[29] Yaochen Liu, Yazhou Zhang, and Dawei Song. 2023. A quantum probability driven framework for joint multi-modal sarcasm, sentiment and emotion analysis. *IEEE Transactions on Affective Computing* 15, 1 (2023), 326–341.

[30] Alejandro López-Cifuentes, Marcos Escudero-Vinolo, Jesús Bescós, and Álvaro García-Martín. 2020. Semantic-aware scene recognition. *Pattern Recognition* 102 (2020), 107256.

[31] Mark Marsden, Eva Mohedano, Kevin McGuinness, Andrea Calafell, Xavier Giró-i Nieto, Noel E O'Connor, Jiang Zhou, Lucas Azevedo, Tobias Daudert, Brian Davis, et al. 2016. Dublin City University and partners' participation in the INS and VTT tracks at TRECVid 2016. In *TRECVID*.

[32] Robert McKee. 1997. *Story: style, structure, substance, and the principles of screenwriting*. Harper Collins.

[33] Majid Mohammadi and Jafar Rezaei. 2020. Ensemble ranking: Aggregation of rankings produced by different multi-criteria decision-making methods. *Omega* 96 (2020), 102254.

[34] Yanrui Niu, Chao Liang, Ankang Lu, Baojin Huang, Zhongyuan Wang, and Jiahao Guo. 2023. Person-action instance search in story videos: An experimental study. *ACM Transactions on Information Systems* 42, 2 (2023), 1–34.

[35] Arpan Phukan and Asif Ekbal. 2023. QeMMA: Quantum-Enhanced Multi-Modal Sentiment Analysis. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, Jyoti D. Pawar and Sobha Lalitha Devi (Eds.). NLP Association of India (NLPAI), Goa University, Goa, India, 815–821. https://aclanthology.org/2023.icon-1.84/

[36] Arpan Phukan, Anas Anwarul Haq Khan, and Asif Ekbal. 2024. QuMIN: quantum multi-modal data fusion for humor detection. *Multimedia Tools and Applications* (2024), 1–18.

[37] Arpan Phukan, Santanu Pal, and Asif Ekbal. 2024. Hybrid Quantum-Classical Neural Network for Multimodal Multitask Sarcasm, Emotion, and Sentiment Analysis. *IEEE Transactions on Computational Social Systems* 11, 5 (2024), 5740–5750. doi:10.1109/TCSS.2024.3388016

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[39] Benjamin Reilly. 2002. Social choice in the south seas: Electoral innovation and the borda count in the pacific island countries. *International Political Science Review* 23, 4 (2002), 355–372.

[40] Prayag Tiwari, Lailei Zhang, Zhiguo Qu, and Ghulam Muhammad. 2024. Quantum fuzzy neural network for multimodal sentiment and sarcasm detection. *Information Fusion* 103 (2024), 102085.

[41] Zhengqian Wu, Ruizhe Li, Zijun Xu, Zhongyuan Wang, Chunxia Xiao, and Chao Liang. 2025. FriendsQA: A New Large-Scale Deep Video Understanding Dataset with Fine-grained Topic Categorization for Story Videos. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence*.

[42] Jian Xiao, Zhenzhen Hu, Jia Li, and Richang Hong. 2024. Text Proxy: Decomposing Retrieval from a 1-to-N Relationship into N 1-to-1 Relationships for Text-Video Retrieval. *arXiv preprint arXiv:2410.06618* (2024).

[43] Yu Xiao, Hong-Zhong Deng, Xin Lu, and Jun Wu. 2021. Graph-based rank aggregation method for high-dimensional and partial rankings. *Journal of the Operational Research Society* 72, 1 (2021), 227–236.

[44] William Zeng and Bob Coecke. 2016. Quantum algorithms for compositional natural language processing. *arXiv preprint arXiv:1608.01406* (2016).

[45] Xing Zhang, Zuxuan Wu, and Yu-Gang Jiang. 2021. SAM: Modeling scene, object and action with semantics attention modules for video recognition. *IEEE Transactions on Multimedia* 24 (2021), 313–322.

[46] Yazhou Zhang, Dawei Song, Peng Zhang, Panpan Wang, Jingfei Li, Xiang Li, and Benyou Wang. 2018. A quantum-inspired multimodal sentiment analysis framework. *Theoretical Computer Science* 752 (2018), 21–40. doi:10.1016/j.tcs.2018.04.029 Quantum structures in computer science: language, semantics, retrieval.