Table 7. GamePoint@K comparison.

Method	GamePoint@K-1	GamePoint@K-5	GamePoint@K-10	GamePoint@K-50	GamePoint@K-100
Qwen2.5-VL-7B	0.330	0.328	0.331	0.330	0.329
SWIM	**0.373**	**0.375**	**0.374**	**0.373**	**0.374**

Appendix

A. Benchmarks

For completeness, we provide detailed descriptions of the general benchmarks used in the main paper.

ActivityNet-QA [3] is a large-scale video question answering benchmark constructed from the ActivityNet dataset. It contains human-annotated question–answer pairs focusing on action-related content, with an average video duration of about 2 minutes. The questions are designed to require understanding of dynamic scenes and temporal sequences rather than static visual cues.

VideoMME [20] collects videos from a wide range of domains, including sports, documentaries, instructional content, and entertainment. Video durations vary from minutes to hours, making it one of the most comprehensive and challenging benchmarks for holistic video understanding. The diversity in topic, style, and duration tests a model's ability to handle long-context reasoning and adapt to varied visual–text scenarios.

MVBench [35] is a multi-choice video understanding benchmark comprising 20 distinct tasks. Each task presents a multiple-choice question targeting temporal comprehension, covering scenarios such as event ordering, cause–effect reasoning, motion tracking, and activity prediction. These tasks require sophisticated temporal reasoning and understanding of dynamic content that cannot be solved by analyzing a single frame, thereby evaluating a model's capability to integrate information across time.

Together, these benchmarks provide a diverse evaluation landscape: ActivityNet-QA and VideoMME emphasize broad video understanding with varying domain coverage and length, whereas MVBench focuses on fine-grained temporal reasoning across multiple types of challenges.

B. More Experimental Analysis

B.1. GamePoint@K

We further evaluate retrieval accuracy using GamePoint@K, which measures the fraction of relevant elements among the top-K highest-scoring positions in the attention map:

$$\text{GamePoint@K} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\text{TopK}(\bar{\mathbf{A}}_i, K) \cap P_i|}{|\text{TopK}(\bar{\mathbf{A}}_i, K)|}, \quad (9)$$

where $\text{TopK}(\bar{\mathbf{A}}_i, K)$ selects the K highest-scoring elements for sample i, and P_i denotes its ground-truth positions. Higher GamePoint@K scores indicate that relevant visual targets are ranked closer to the top, reflecting better alignment between textual references and visual regions.

As shown in Table 7, SWIM consistently outperforms Qwen2.5-VL across all K values. At $K = 1$, SWIM achieves **0.373** compared to 0.330 for Qwen2.5-VL, indicating stronger ability to position the correct target at the top rank. This advantage is maintained at broader retrieval depths, with SWIM reaching **0.375** at $K = 5$ (+4.7% over baseline) and retaining stable gains for $K = 10$, $K = 50$, and $K = 100$. The consistent margins across different K suggest that SWIM produces reliable ranking distributions, keeping relevant objects prominent even as the retrieval list expands.

B.2. Robustness to Synonym-based Linguistic Noise

To assess the robustness of SWIM to variations in referring expressions, we conduct an evaluation in which words enclosed in `<ins>` tags within the VideoRefer-Bench-D prompts are replaced by semantically equivalent synonyms. This modification leaves the overall meaning unchanged but alters the surface form of the text, introducing lexical noise that may challenge models relying on exact token matches. Such a setting reflects real-world scenarios where object references may vary due to differences in speaker style, domain-specific terminology, or translation artifacts, and tests whether a model can preserve grounding accuracy under these conditions. As shown in Table 8, the original SWIM achieves an average score of 3.78, while SWIM* obtains 3.74 under synonym perturbations—a marginal drop of 0.04. Compared to Qwen2.5-VL, SWIM maintains strong performance under synonym substitutions, achieving an average accuracy of **3.74** against 3.43. These results indicate that SWIM's alignment mechanism is resilient to changes in word choice, preserving its ability to ground natural language object references to the correct visual regions even under lexical variation.

Table 8. Performance comparisons on VideoRefer-Bench-D. * denotes incorporating synonym-based noise.

	VideoRefer-Bench-D				
Method	**SC**	**AD**	**TD**	**HD**	**Avg.**
Qwen2.5-VL-7B [2]	3.99	3.05	2.44	2.44	2.97
Qwen2.5-VL-7B* [2]	4.78	3.49	3.27	2.18	3.43
SWIM	**4.92**	**3.85**	**3.43**	2.96	**3.78**
SWIM*	**4.86**	**3.78**	**3.36**	2.96	**3.74**