Learning to Refuse: Refusal-Aware Reinforcement Fine-Tuning for Hard-Irrelevant Queries in Video Temporal Grounding

Supplementary Material

A. Metrics

We provide detailed descriptions of the evaluation metrics used in the main paper, including RA-IoU, F1 score, RT-IoU, Sentence-BERT score, and LLM-based score.

RA-IoU. RA-IoU is an mIoU-based metric that incorporates relevance classification, following prior work [6]. (1) If the query is relevant and the model outputs a timestamp, we compute the IoU between the predicted segment and the ground-truth segment. (2) If the query is irrelevant and the model does not produce a timestamp, we assign a score of 1. (3) Otherwise, we assign a score of 0. R@m denotes the proportion of samples whose RA-IoU is greater than a threshold m.

RT-IoU. RT-IoU measures the alignment between the semantic relevance category mentioned in a refusal answer and the ground-truth category used to construct its corresponding hard-irrelevant query. When extracting the categories from the refusal answer, we provide the generated refusal answer and the definitions of the semantic categories as input to GPT-5-mini. Using the extracted categories and the ground-truth categories, RT-IoU is computed as the intersection of categories divided by the union of categories.

SBert Score. We compute the cosine similarity between the Sentence-BERT [34] embeddings of the generated refusal answer and the ground-truth refusal answer, producing a score in the range of 0 to 1.

LLM Score. We use GPT-5-mini [1] to assess the semantic consistency between the generated refusal answer and the ground-truth refusal answer, with the LLM producing a consistency score in the range of 1 to 5.

B. Semantic Relevance Category Definition

To construct hard-irrelevant queries, we define eleven semantic relevance categories grouped into four high-level types. These categories describe the possible relationships between a video and a text query, and are defined to reflect the spatiotemporal characteristics of the video. Fig. 5 provides detailed descriptions of each semantic relevance category.

C. Experimental Results via Category Types

To analyze relevance discrimination across semantic relevance categories, we evaluate performance using the categories employed to construct the hard-irrelevant queries. Figure 6 shows that our model outperforms the baseline across all categories. The model achieves strong improvements across categories requiring spatial understanding, such as Object Existence, Scene Existence, and Attribute Value, as well as categories involving temporal understanding, including Action Sequence, Object Moving, and Scene Transition. These results indicate that our approach effectively enhances relevance discrimination across a wide range of semantic mismatch types.

High-level	Low-level	Definition
Action	Action Sequence	Two or more distinct actions occur sequentially.
	Fine-grained Action	A single action that is visually similar to another but distinguishable through fine temporal or motion cues.
Object	Object Existence	Distinct, identifiable objects are visually present.
	Object Part Relation	Two or more objects exhibit a whole–part relationship, requiring compositional understanding of object structures.
	Object Spatial Relation	Objects maintain explicit spatial or directional relationships, such as position or orientation.
	Object Moving	An object moves with a visible trajectory or direction
Scene	Scene Existence	A clearly distinguishable environment or background appears.
	Scene Transition	Two or more distinct scenes appear sequentially, showing a transition or change over time.
Attribute	Attribute Value	Objects or agents exhibit intrinsic visual properties such as color, size, or emotion.
	Counting	The number of entities or actions is visually identifiable in a scene.
	Comparison	Two or more entities appear with explicit comparative attributes.

Figure 5. Semantic Relevance Category definition

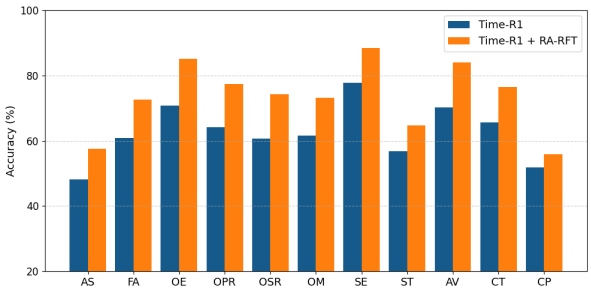

Figure 6. Performance analysis via semantic relevance categories.

D. Comparison of Supervised Fine-Tuning Method

To analyze the effectiveness of the proposed RA-RFT strategy, we compare our method with supervised fine-tuning (SFT). Table 8 shows the results of training on our HI-VTG dataset using SFT. The SFT-trained model shows reduced performance across all metrics, likely due to catastrophic forgetting caused by imitating instruction-formatted answers. In contrast, the model trained with RA-RFT achieves higher RA-IoU and F1 scores, demonstrating the effectiveness of our approach.

Method	RA-IoU				F1-Score		
	R@.3	R@.5	R@.7	mIoU	rel.	irrel.	avg.
Time-R1	53.0	45.0	38.2	45.9	74.2	66.9	70.5
+SFT	29.5	20.4	11.8	20.9	0.1	0.0	0.0
+RA-RFT (ours)	59.6	51.3	43.8	51.9	77.6	75.0	76.3

Table 8. Comparison with other fine-tuning methods.

Type	Method	RA-ActivityNet			RA-Charades			RA-TVGBench		
		RT-IoU	SBert score	LLM score	RT-IoU	SBert score	LLM score	RT-IoU	SBert score	LLM score
SFT	TimeChat [35]	0.0	0.00	0.00	0.0	0.00	0.00	0.0	0.00	0.00
	TimeSuite [46]	0.0	0.00	0.00	0.0	0.00	0.00	0.0	0.00	0.00
	TRACE [14]	0.0	0.00	0.01	0.0	0.00	0.00	0.0	0.02	0.00
RFT	Time-R1 [24]	30.4	0.41	2.00	31.7	0.40	1.76	24.3	0.38	1.51
	+ RA-RFT (ours)	37.1	0.51	2.44	39.1	0.49	2.10	33.7	0.55	2.18
	VideoChat-R1 [22]	25.9	0.35	1.72	25.3	0.32	1.44	18.4	0.29	1.28
	+ RA-RFT (ours)	34.4	0.48	2.32	40.7	0.49	2.33	26.6	0.41	1.85
	VideoChat-R1-think [22]	24.6	0.34	1.67	20.9	0.28	1.28	27.8	0.42	1.88
	+ RA-RFT (ours)	34.3	0.48	2.32	37.7	0.46	2.23	30.6	0.45	1.99

Table 9. Refusal explanation quality on the RA-VTG evaluation datasets.

E. Refusal Explanation Quality on the HI-VTG Dataset

Table 9 presents the refusal explanation quality on the three RA-VTG evaluation datasets. Models trained with RA-RFT achieve higher scores across RT-IoU, SBERT score, and LLM-based score compared to the base models. These results indicate that our method improves the model's ability to generate clearer and more appropriate refusal explanations for hard-irrelevant queries across diverse evaluation settings.

F. More Experimental Results via Difficulty

To complement the main paper's analysis in Sec. 5.2, we further evaluate RA-RFT across different levels of hard-irrelevance on other datasets. Table 10 and Tab. 11 show the HI-Charades and HI-VTGBench results. RA-RFT consistently improves both refusal accuracy and explanation quality at all difficulty levels, with the largest gains appearing in the strong hard-irrelevant setting where fine-grained reasoning is crucial. Overall, these extended results across multiple datasets show that RA-RFT generalizes well across different levels of semantic discrepancy and consistently improves both refusal ability and explanation quality in diverse hard-irrelevant VTG scenarios.

Method	Strong Hard.		Moderate Hard.		Weak Hard.	
	F1	LLM sc.	F1	LLM sc.	F1	LLM sc.
RaTSG	53.6	-	66.3	-	74.5	-
Time-R1	63.2	1.68	73.3	1.76	82.8	1.98
+ RA-RFT (ours)	71.5	1.95	82.1	2.22	90.0	2.29
VideoChat-R1	50.7	1.19	65.8	1.61	73.3	1.77
+ RA-RFT (ours)	70.4	2.09	84.6	2.57	90.3	2.54
VideoChat-R1-think	47.7	1.08	59.3	1.37	72.1	2.48
+ RA-RFT (ours)	69.1	2.00	82.1	2.43	88.5	2.48

Table 10. Performance across different levels of hard-irrelevance on HI-Charades dataset.

G. Performance Analysis of the Qwen2.5-VL-7B Base Model Trained From Scratch

To demonstrate the effectiveness of our HI-VTG training data and RA-RFT strategy on a general LVLM, we apply our method to the Qwen2.5-VL-7B model, which is not originally trained for the VTG task. The model is trained for 3 epochs. As shown in Tab. 12, with our method incorporated, the model achieves higher performance in both VTG grounding and relevance discrimination. This indicates that the our dataset and learning strategy are effective in enabling the model to refuse irrelevant queries and perform temporal grounding.

Method	Strong Hard.		Moderate Hard.		Weak Hard.	
	F1	LLM sc.	F1	LLM sc.	F1	LLM sc.
Time-R1	52.5	0.95	66.2	1.49	78.8	2.08
+ RA-RFT (ours)	67.5	1.57	84.7	2.20	91.6	2.77
VideoChat-R1	58.0	0.80	74.6	1.29	80.7	1.74
+ RA-RFT (ours)	64.1	1.41	76.1	1.82	85.6	2.30
VideoChat-R1-think	62.4	1.36	76.4	2.05	77.1	2.23
+ RA-RFT (ours)	70.1	1.62	81.6	2.03	85.6	2.33

Table 11. Performance across different levels of hard-irrelevance on HI-TVGBench dataset.

Method	RA-IoU				F1-Score		
	R@.3	R@.5	R@.7	mIoU	rel.	irrel.	avg.
Qwen2.5-VL-7B	32.9	26.2	21.2	27.8	69.8	41.2	55.5
+RA-RFT (ours)	55.4	48.6	42.9	49.6	75.2	73.1	74.2

Table 12. Performance of Qwen2.5-VL-7B trained from scratch.

H. Evaluation on Refusal-aware Video Question Answering

To examine whether RA-RFT generalizes beyond the VTG task, we conduct an additional experiment on refusal-aware video question answering (Video QA) using the UVQA dataset. In this setting, the model answers questions when they are relevant to the video and refuses when they are irrelevant. As shown in Tab. 13, RA-RFT improves overall performance on both relevant and irrelevant queries. These results suggest that the proposed RA-RFT generalizes beyond VTG to broader video understanding tasks.

Method	Rel. query				Irrel. query		
	Pre.	Rec.	F1	Answer Acc.	Pre.	Rec.	F1
Time-R1	92.8	86.0	89.3	24.7	87.0	93.3	90.0
+ RA-RFT (ours)	91.9	90.7	91.3	26.7	90.8	92.0	91.4

Table 13. Results on Refusal-aware Video QA dataset

I. Standard VTG Performance with Additional Baselines

We further evaluate standard VTG performance on two representative benchmarks, ActivityNet Captions and Charades-STA. As shown in Tab. 14, our method preserves the grounding capability of Time-R1 and achieves competitive performance compared to learning-based VTG methods, even in a zero-shot setting.

Method	ActivityNet Captions			Charades-STA		
	R@.5	R@.7	mIoU	R@.5	R@.7	mIoU
2D-TAN*	43.4	25.0	42.5	43.4	25.0	41.1
VSLNet*	43.2	26.2	43.2	42.7	24.1	41.6
Moment-DETR*	-	-	-	52.1	30.6	45.5
QD-DETR*	-	-	-	-	32.6	-
RaTSG*	38.8	22.5	39.8	41.0	26.1	38.3
Time-R1	40.1	22.0	41.2	62.0	35.7	54.0
Time-R1+ RA-RFT	40.2	21.7	41.3	61.8	35.2	53.6

Table 14. VTG Performance comparison with representative baselines on standard VTG benchmarks. * indicates the model trained and evaluated under the same data distribution

J. Qualitative Results

Figure 7 and Fig. 8 present qualitative results for strong hard-irrelevant queries from HI-ActivityNet. Also, Figure 9 shows qualitative results for hard-irrelevant queries from the human-annotated RA-VTG dataset. While previous methods often fail to refuse hard-irrelevant queries and predict temporal segments, the models trained with RA-RFT effectively refuse these hard-irrelevant queries. In addition, our models clearly explain the refusal reasons and correctly reconstruct the original queries.

K. Prompt Details

We used GPT-5-mini to construct the HI-VTG dataset. Table 15 shows the prompts used for extracting semantic relevance categories from a given video and query. Also, Tab. 16 shows the prompts used for generating hard-irrelevant queries and corresponding refusal answers. To evaluate the explanation quality of the model outputs, we used the prompts in Table 17 and Tab. 18, and extracted RT-IoU and LLM-based scores.

Video:

Query: A chef is shown standing in front of a plain wall.

Ground Truth:

<think><timestep>0.0 to 5.2</timestep> A chef is shown standing in front of a food related mural on the wall.<timestep>5.2 to 103.2</timestep> Then various ingredients are shown and he goes through the process of making some sandwich rolls with tortilla.<timestep>40.8 to 103.2</timestep> He lays on the different ingredients and then rolls them all up in the tortilla, cutting them into slices afterwards.</think><answer>There is a noticeable mismatch: the query does not align with the video segment. The query states the chef stands in front of a plain wall, but the opening shot clearly shows a food-related mural on the wall, so no part of the footage depicts a plain wall in front of the chef.</answer>

VideoChat-R1:

<think>The video starts with Chef Herman standing in front of a plain wall, which matches the query. The subsequent frames show the preparation of a wrap, but they do not include the chef standing in front of a plain wall again.</think><answer>0.00 to 5.00</answer>

VideoChat-R1 + RA-RFT:

<think>The query states that the video should show a chef standing in front of a plain wall. However, the video starts with Chef Herman from Damascus Bakeries standing in front of a colorful backdrop featuring images of food and bakery products. This does not match the description of a plain wall.</think><answer>The video does not contain a relevant segment because it starts with Chef Herman standing in front of a colorful backdrop, which contradicts the requirement for a plain wall.</answer><correction>A chef is shown standing in front of a colorful backdrop.</correction>

VideoChat-R1-think:

<think>The video starts with Chef Herman standing in front of a plain wall, which matches the given query. However, as the video progresses, it shows the preparation of a wrap sandwich, which is not related to the initial scene of the chef standing in front of a plain wall. Therefore, the relevant segment for the query ends at the point where the chef stops standing in front of the plain wall.</think><answer>0.00 to 6.00</answer>

VideoChat-R1-think + RA-RFT:

<think>The query states that the chef is standing in front of a plain wall. However, the video shows the chef standing in front of a backdrop with various food advertisements. Therefore, there is no segment in the video that matches the description of the chef standing in front of a plain wall.</think><answer>The video does not contain a segment that is relevant to the query. The chef is standing in front of a backdrop with various food advertisements, not a plain wall.</answer><correction>A chef is shown standing in front of a backdrop with various food advertisements.</correction>

Figure 7. Qualitative results for strong hard-irrelevant queries from HI-ActivityNet.

Video:

Query: A man is seen from different angles as he continues rowing.

Ground Truth:

<think><timestep>0.0 to 19.5</timestep> a man is sailing on a boat in the ocean.<timestep>26.9 to 90.5</timestep> He is seen going through the rough waters with his sail flying in the wind.<timestep>95.7 to 114.5</timestep> He is seen from different angles as he continues sailing.</think>answer>The query does not fully align with the video segment. The segment shows him sailing on a boat in the ocean, while the query describes him rowing; no portion depicts rowing.</answer>

Time-R1:

<think>The query mentions that the person is seen from different angles as they continue rowing. The video shows the person sailing a boat through rough waters, and it does show the person from different angles, indicating that the person is indeed continuing to sail.</think><answer>The relevant segment is from 10.00 to 100.00 seconds.</answer>

Time-R1 + RA-RFT:

<think>The query "He is seen from different angles as he continues rowing." is not directly related to the video content. The video primarily shows a sailboat navigating through rough seas, and the person on the boat is not rowing but rather sailing.</think><answer>The video does not contain a relevant segment because it does not show a person rowing a boat. Instead, it shows a sailboat navigating through rough seas.</answer><correction>"He is seen from different angles as he continues sailing."</correction>

VideoChat-R1-think:

<think>The video shows a sailboat navigating through rough seas. The sailboat is seen from various angles as it maneuvers through the waves. The sailboat's movements and the changing perspectives suggest that the sailboat is being rowed or steered by someone on board. Therefore, the video contains a segment that is relevant to the query.</think><answer>6.00 to 11.00</answer>

VideoChat-R1-think + RA-RFT:

<think>The query "He is seen from different angles as he continues rowing." suggests that there should be a segment where a person is shown rowing a boat from various perspectives. However, the video shows a sailboat navigating through rough seas, which indicates that the person is sailing rather than rowing. Therefore, the video does not contain a relevant segment for the given query.</think><answer>The video does not contain a relevant segment for the given query because it shows a sailboat navigating through rough seas, indicating that the person is sailing rather than rowing.</answer><correction>He is seen from different angles as he continues sailing.</correction>

Figure 8. Additional qualitative results for strong hard-irrelevant queries from HI-ActivityNet.

Video:

Query: A man is cleaning a red car with a blue towel.

Ground Truth:

<answer>This query does not align well with the video. The color of the towel the man uses to wipe the car is orange, not blue.</answer>

Time-R1:

<think>The query “A man is cleaning a red car with a blue towel." is relevant to the video segment from 17.83 to 26.67 seconds. In this segment, the man is seen using a blue towel to wipe down the car, which aligns with the query.</think><answer>The segment from 17.83 to 26.67 seconds is relevant to the query.</answer>

Time-R1 + RA-RFT:

<think>The query “A man is cleaning a red car with a blue towel." is not directly relevant to the video content. The video shows a man cleaning a red car with a cloth, but the cloth is orange, not blue.</think><answer>The video does not contain a relevant segment for the query “A man is cleaning a red car with a blue towel." because the cloth used is orange, not blue.</answer><correction>A man is cleaning a red car with an orange towel.</correction>

VideoChat-R1:

<think>The video shows a man cleaning a red car. Initially, he uses a spray bottle to apply a substance onto the car's hood. Then, he switches to a cloth to wipe down the car. The man is wearing an orange shirt and blue jeans. The car is parked in front of a garage door. The sequence of actions matches the query closely, as the man is indeed cleaning the car with a cloth.</think><answer>12.00 to 25.00</answer>

VideoChat-R1 + RA-RFT:

<think>The video shows a man cleaning an orange car. The query mentions a "red car" which seems to be a mismatch as the car in the video is orange. However, the action described in the query is relevant as the man is indeed cleaning the car with a towel. The towel mentioned in the query could be a misinterpretation of the orange cloth used by the man in the video.</think><answer> The video does not contain a relevant segment for the query because the man is cleaning an orange car with an orange towel, not a red car, blue towel.</answer><correction>A man is cleaning an orange car with an orange towel.</correction>

Figure 9. Qualitative results for hard-irrelevant queries from human-annotated RA-VTG dataset.

[SYSTEM] You are a multi-label classifier for building Video Temporal Grounding (VTG) datasets.

Task introduction
- The input "related_query" is a valid VTG query.
- Do NOT rewrite the query.
- Identify category paths that can modify the query into a plausible but video-irrelevant (negative) query.
- Select all applicable category paths and briefly justify why each can produce an irrelevant query.

Category taxonomy (Parent / Child)

Action
- Action/ActionSequence — Change the temporal ordering of actions.
- Action/FineGrainedAction — Replace an action verb with a visually similar but directionally or temporally distinct one.

Object
- Object/ObjectExistence — Add or remove an identifiable object.
- Object/ObjectPartRelation — Modify part–whole relations or accessory relations.
- Object/ObjectSpatialRelation — Change relative spatial positions of objects.
- Object/ObjectMoving — Change the motion direction or trajectory of an object.

Scene
- Scene/SceneExistence — Replace the type of scene.
- Scene/SceneTransition — Change scene order, transition direction, or timing.

Attribute
- Attribute/AttributeValue — Change intrinsic properties such as color, size, material, shape, or state.
- Attribute/Counting — Change the number of objects or actions.
- Attribute/Comparison — Flip comparative relations such as size or speed.

Output requirements
- Output ONLY the JSON below.
- Use exact "Parent/Child" names.
- Return all applicable, distinct categories, sorted by diagnostic strength.
- Output at least 3 categories whenever possible.

[OUTPUT JSON]
{
  "eligible_categories": [
    {"path": "Parent/Child", "reason": "<justification>"},
    {"path": "Parent/Child", "reason": "<justification>"}
  ]
}

Table 15. Prompt for extracting semantic relevance categories from a given query using an LLM.

[SYSTEM] You generate hard negative (irrelevant) queries for Video Temporal Grounding (VTG).

Purpose
- Create irrelevant (negative) queries from valid related queries.
- For each negative query, generate structured reasoning that a Video-LLM could output.

Task
- Input: related_query, reference timestamp, optional video_context, and category-based plans.
- For each plan:
  1. Edit the related_query ONLY along its categories to produce one irrelevant query.
  2. Generate reasoning using the REQUIRED block format:
     <irrelevant_answer>...</irrelevant_answer><category1>...</category1><category2>...</category2>...

- <irrelevant_answer>block:
  * Must state misalignment between the query and the video.
  * Strength depends on difficulty (high/medium/easy).

- <category>blocks:
  * One block per applied category (tag = path lowercased with slashes as underscores).
  * Briefly explain why the irrelevant query does not match the video for that category.

- Difficulty levels:
  - 1 category → strong
  - 2 categories → moderated
  - 3 categories → weak

Category taxonomy (Parent / Child)
(Same category taxonomy as in the semantic relevance category extraction prompt.)

Input format
- related_query: "<string>"
- related_query_timestamp: "<start>-<end>second"
- plans: list of 1–3 items, each with difficulty and applied_categories.
- video_context:
  * A textual description of the video, provided as a string OR a JSON array of strings.
  * Each entry may include an associated time range and a natural-language description of what appears in the video.

Output format (JSON only)
The model must output a single JSON object. Each difficulty present in the input plans must appear once under "negs". All fields must be included exactly as shown.

```
{
  "negs": {
    "<difficulty>": {
      "irrel_query": "<generated negative query>",          % negative query produced via category edits
      "applied_categories": [
        {"path": "Parent/Child"}, ...                       % same paths & order as in the plan
      ],
      "reasoning":
        "<irrelevant_answer>... </irrelevant_answer>"
        "<parent_child>... </parent_child>" ... ,           % one block per applied category
      "difficulty_tag": "<difficulty>"                      % must match the plan
    }
  }
}
```

Table 16. Prompt for generating hard-irrelevant queries and refusal answers. A given query, extracted semantic relevance categories, and video context are used to generate an irrelevant query using an LLM.

[SYSTEM]
You are a strict multi-label classifier.
Identify which reasoning categories from the video–text mismatch categories below are used or implied in a Generated Response that explains why a query is irrelevant to a video.
Select all applicable categories according to the meaning expressed in the response.

Video–Text Mismatch Categories
(Same category taxonomy as in the semantic relevance category extraction prompt.)

Rules
1. Include a category only if it is clearly supported or implied by the reasoning.
2. Multiple categories may apply, but avoid redundant or speculative labels.
3. Use only the exact category paths listed above.
4. Ignore style, tone, or fluency — focus purely on reasoning content.
5. If none apply, return an empty list.

Output Format
Return only a JSON array of strings containing the selected categories.
Examples:
[“Object/ObjectExistence”, “Attribute/Counting”]
If none apply: []

Table 17. Prompt for evaluating RT-IoU between the semantic categories in the model’s refusal answer and the ground-truth categories.

[SYSTEM]
You are a evaluator designed to assess the reasoning consistency between a Generated Response and a Ground Truth (GT) Response.

TASK:
Your job is to evaluate how faithfully the Generated Response reproduces the reasoning in the GT Response.

INSTRUCTIONS:
Reasoning Consistency Evaluation:
- Evaluate how faithfully the Generated Response reproduces the reasoning and justification in the GT Response.
- A consistent response must keep the GT’s mismatch points, evidence, and contextual explanations. It must not distort their meaning.
- Omissions or contradictions of GT reasoning elements must be penalized.
- Extra explanations are allowed if they are logically consistent with the GT.
- The reasoning must remain factually and logically compatible with the GT Response.
- Do not consider fluency, tone, or paraphrasing style. Focus only on semantic and factual consistency.

Scoring Scale (0–5):
Assign a score between 0 and 5, allowing decimal values, based on how well the reasoning aligns with the ground-truth reasoning.

Evaluation Mindset:
- You MUST prioritize factual and logical alignment over stylistic similarity.
- Do NOT penalize harmless elaborations.
- You MUST penalize any omission or contradiction of GT reasoning.
- You MUST NOT assign a score above 4.9 unless reasoning is perfectly consistent.

OUTPUT:
Return ONLY a Python dictionary literal. No explanations.

Examples:
‘score’: 4.0
‘score’: 1.5
‘score’: 3.7

Table 18. Prompt for evaluating an LLM score between a model’s refusal answer and the ground-truth response.