JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

NaQ: Leveraging Narrations as Queries to Supervise Episodic Memory

Santhosh Kumar Ramakrishnan[1]   Ziad Al-Halah[2]   Kristen Grauman[1,3]

Poster session: TUE-PM-245

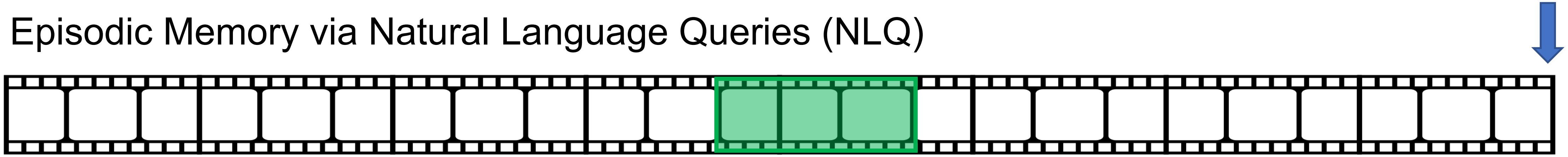
Project page: https://vision.cs.utexas.edu/projects/naq/

Code: https://github.com/srama2512/NaQ

[1] TEXAS The University of Texas at Austin
[2] THE UNIVERSITY OF UTAH
[3] Meta AI

Overview

Task: Episodic Memory via Natural Language Queries (NLQ)



Egocentric video

Response

Query: Did I leave the refrigerator open?

Challenge: Limited training data (e.g., 11k queries over 130 hours of video)

Our idea: Augment NLQ training by learning to localize “narrations”

Example narration text: `C rinses hand; C closes tap`

Convert to NLQ annotation

Narrations-as-query data

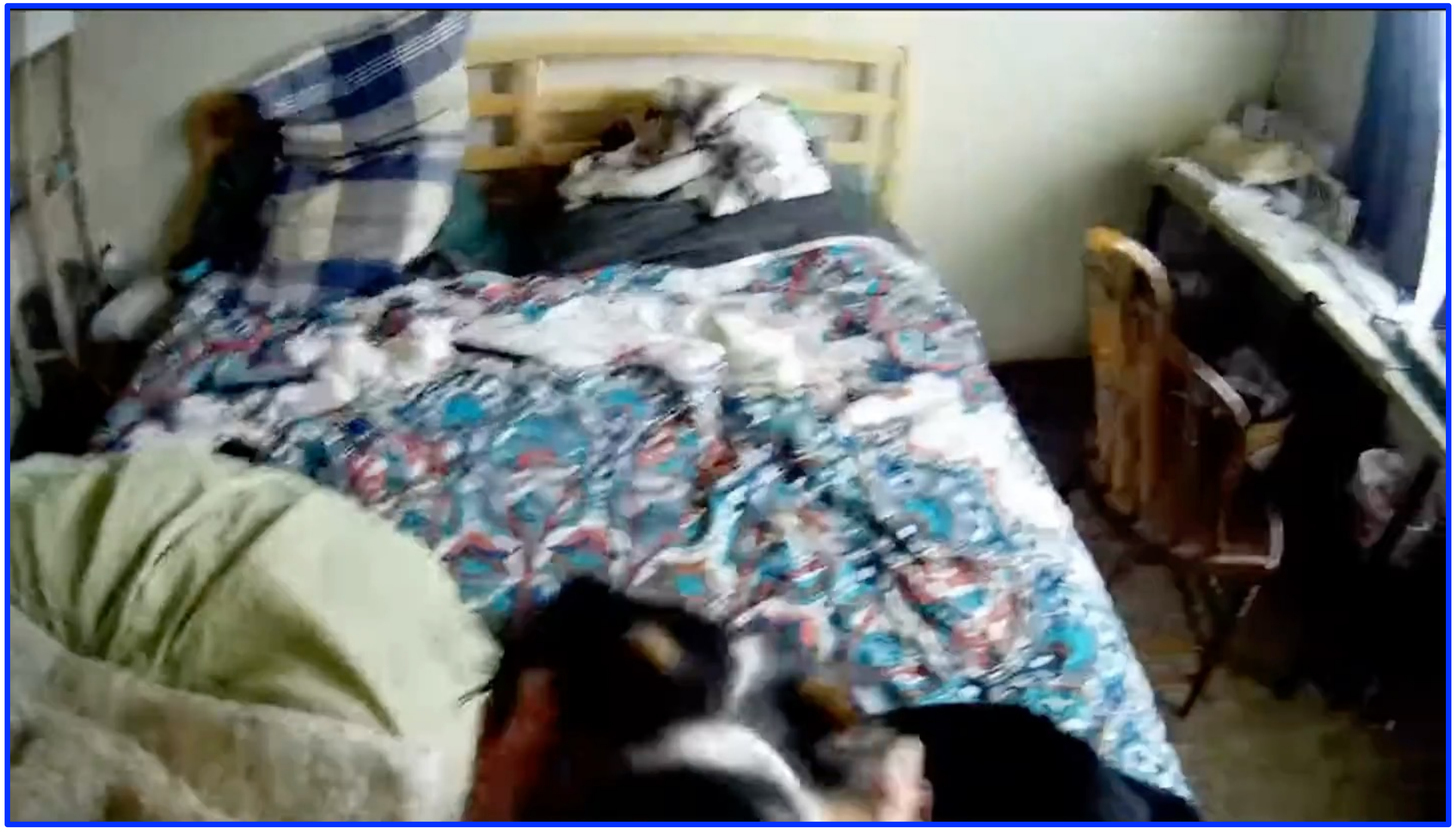
(video, query, response window)

Train NLQ models

Episodic Memory (EM)

Goal: Enable AR assistants for super-human memory

Query: Who did I interact with when I played with the dog for the second time in the living room?

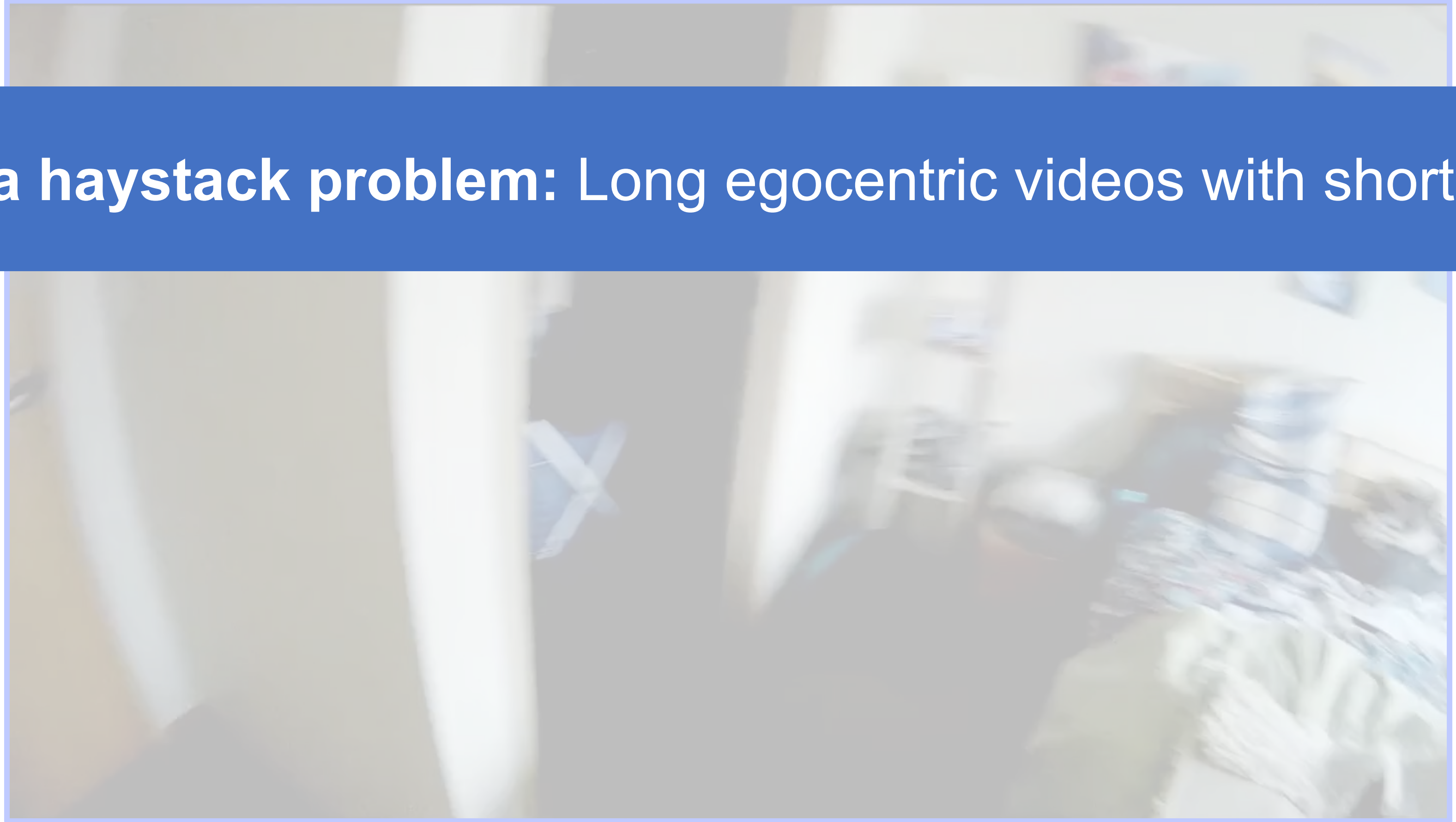


Video credits: Ego4D

Episodic Memory (EM)

Goal: Enable AR assistants for super-human memory

Query: Who did I interact with when I played with the dog for the second time in the living room?



Needle in a haystack problem: Long egocentric videos with short responses

Video credits: Ego4D

Episodic Memory benchmark on Ego4D

Temporally localize responses to **Natural language queries (NLQ)**

Queries formulated based on templates

Category	Template
Objects	Where is object X before / after event Y?
	Where is object X?
	What did I put in X?
	How many X's? (quantity question)
	What X did I Y?
	In what location did I see object X ?
	What X is Y?
	State of an object
	Where is my object X?
Place	Where did I put X?
People	Who did I interact with when I did activity X?
	Who did I talk to in location X?
	When did I interact with person with role X?

NLQ dataset statistics

Split	Train	Val	Test
# video hours	136	45	46
# clips	$1.0k$	$0.3k$	$0.3k$
# queries	$11.3k$	$3.9k$	$4.0k$

Average clip duration: 8.2 mins
Average response duration: 10.5 sec

Needle in a haystack problem

Key challenge: Limited annotation quantity and sparsity

NLQ annotation procedure

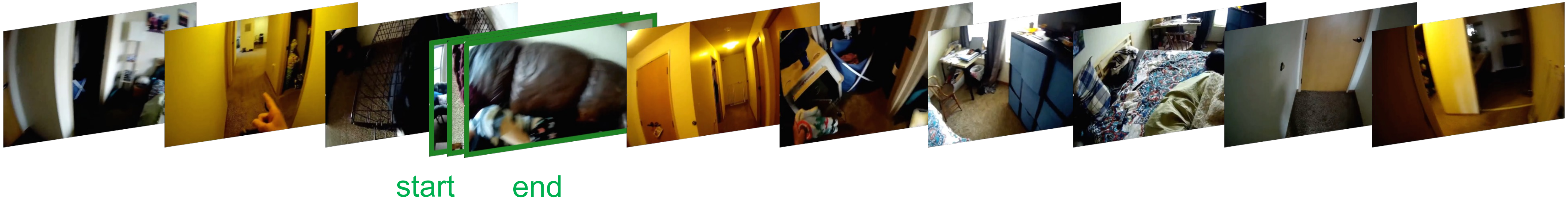
Step 1: Preview long video

Step 2: Formulate creative question

Who did I interact with when I played with the dog for the second time in the living room?

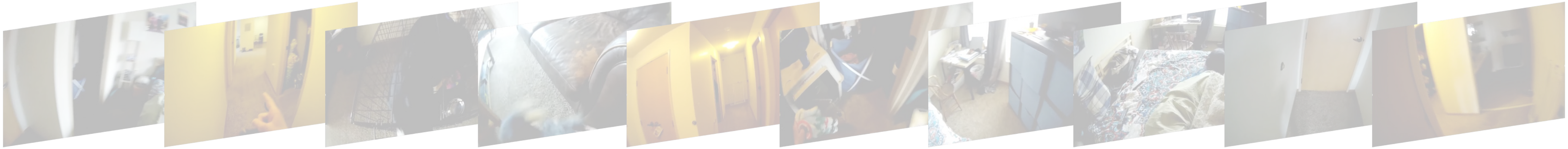
- Template-based
- Unambiguous
- Precise response localization

Step 3: Annotate response (start, end) times



NLQ annotation procedure

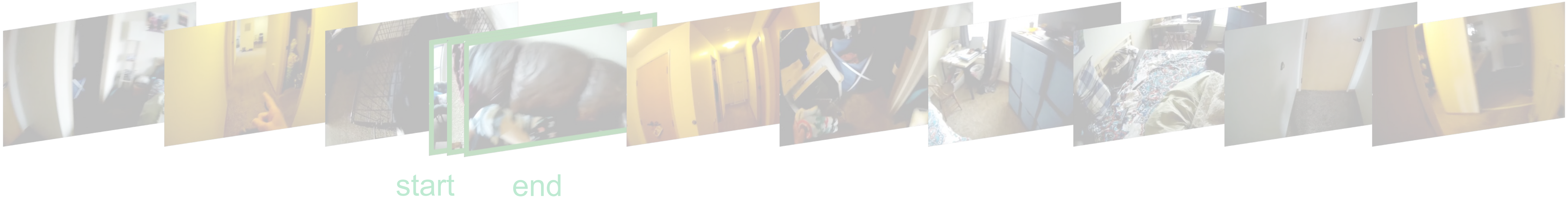
Step 1: Preview long video



Expensive and slow process limits scalability of annotations

- Template-based
- Unambiguous
- Precise response localization

Step 3: Annotate response (start, end) times



NaQ: Narrations-as-Queries

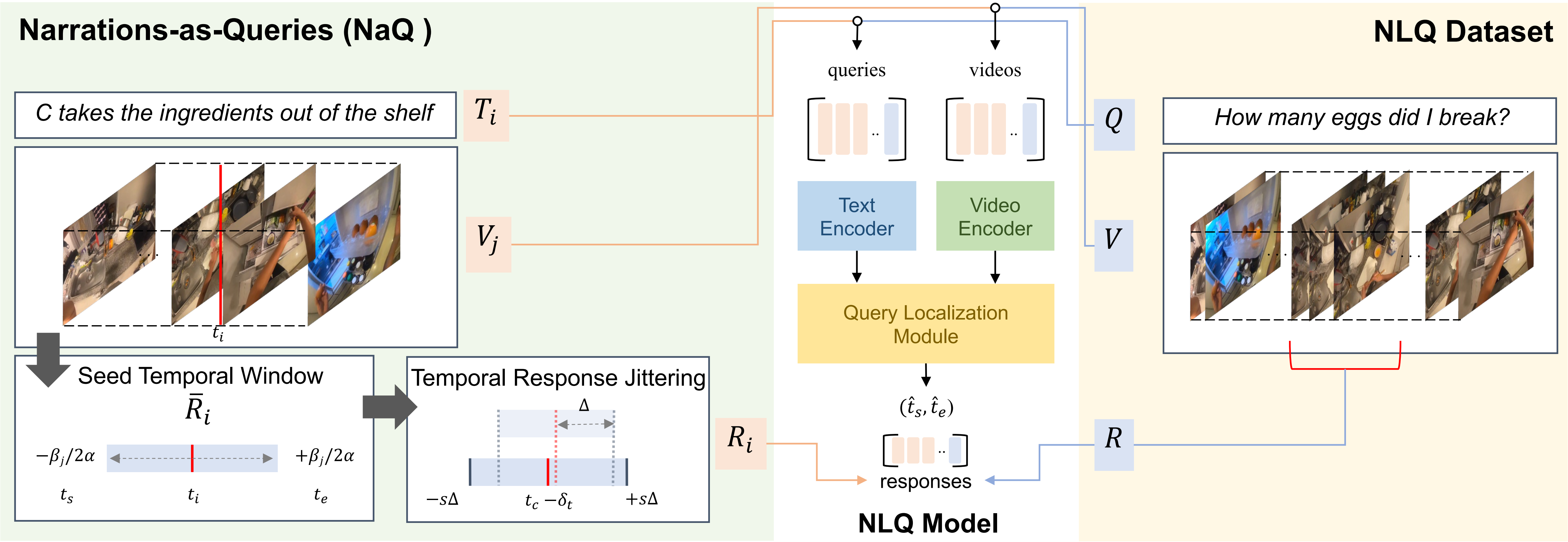
Key insight: Augment NLQ training by learning to localize *narrations*

Timestamped play-by-play descriptions of camera-wearer's activities.

- ✓ **Easier to annotate**
  - Describe as you watch the video
- ✓ **Available on a large scale**
  - 200x more narrations than NLQ annotations
- ✓ **Multi-purpose annotations**
  - Not annotated specifically for NLQ
  - Applications across several benchmarks
  - Likely to be expanded over time

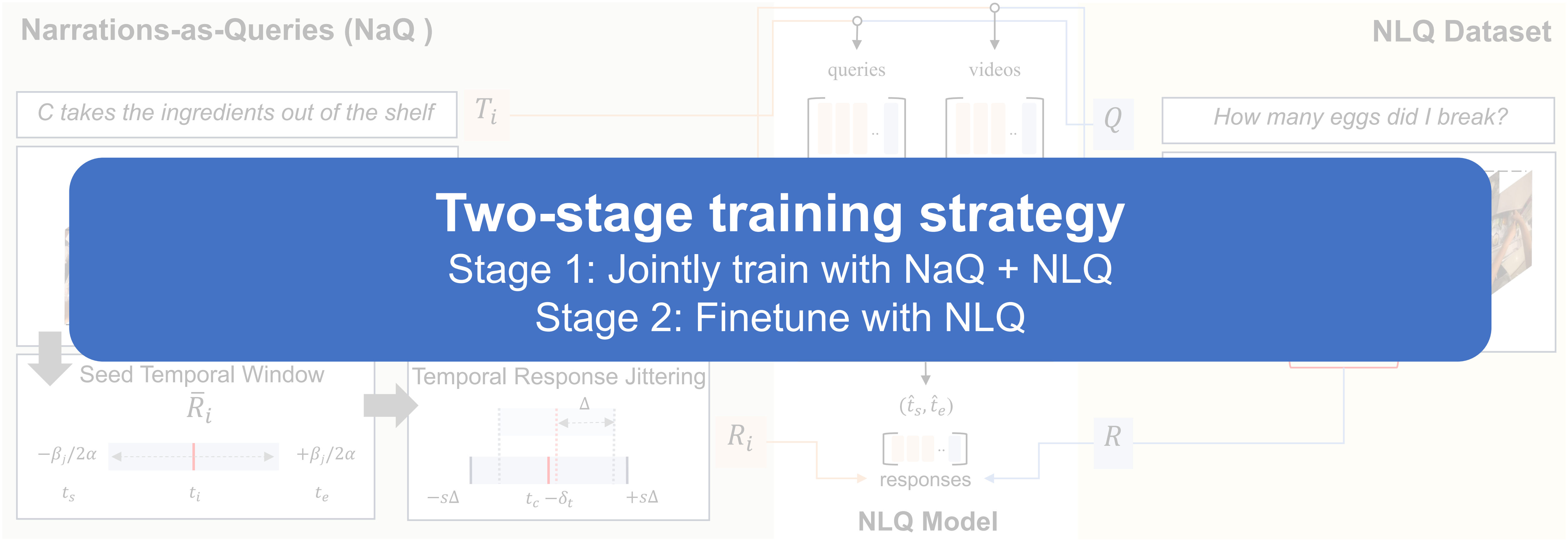
NaQ data-augmentation for scaling NLQ

Simple-yet-effective approach: Augment NLQ dataset using NaQ and perform large-scale training



NaQ data-augmentation for scaling NLQ

Simple-yet-effective approach: Augment NLQ dataset using NaQ and perform large-scale training



Two-stage training strategy

Stage 1: Jointly train with NaQ + NLQ

Stage 2: Finetune with NLQ

Converting narrations → NLQ queries

Narration annotation: $< V_j,\ T_i,\ t_i >$

V_j : Video
T_i : Narration text
t_i : Time-stamp

NaQ annotation for NLQ: $< V_j,\ T_i,\ R_i >$

V_j : Video
T_i : Narration text as query
R_i : (t_s, t_e) response window

T_i *C takes the ingredients out of the shelf*

V_j

t_i

Contextual variable-length clip pairing [1]

Seed Temporal Window

$\bar{R}_i$

$-\beta_j/2\alpha$ $+\beta_j/2\alpha$

t_s t_i t_e

β_j = average separation between consecutive narrations in video j
α = average of β_j over all videos

[1] Lin, Kevin Qinghong, et al. "Egocentric video-language pretraining." *Advances in Neural Information Processing Systems 2022*

Converting narrations → NLQ queries

Narration annotation: $< V_j,\ T_i,\ t_i >$

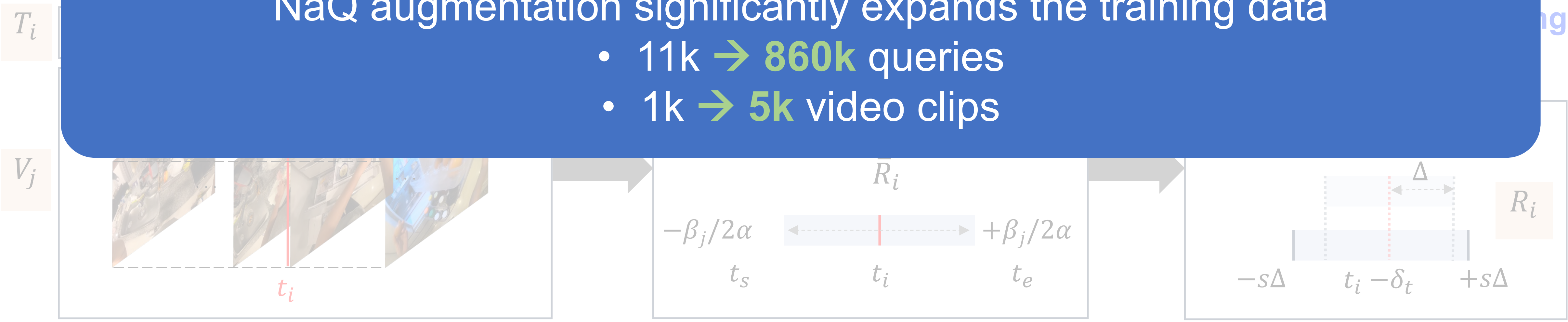
V_j : Video
T_i : Narration text
t_i : Time-stamp

NaQ annotation for NLQ: $< V_j,\ T_i,\ R_i >$

V_j : Video
T_i : Narration text as query
R_i : (t_s, t_e) response window

NaQ augmentation significantly expands the training data

- 11k → **860k** queries
- 1k → **5k** video clips



s = randomly sampled scaling factor
δ_t = random translation factor
Δ = half-width of original temporal window

[1] Lin, Kevin Qinghong, et al. "Egocentric video-language pretraining." *Advances in Neural Information Processing Systems 2022*

Experimental setup

Dataset

Ego4D NLQ dataset [1]

Evaluation metrics

Mean Recall @ k: Recall @ top k retrieval averaged over IoU=[0.3, 0.5]

Baselines

VSLNet [1,2]: Span-based localization approach to vision-language grounding

EgoVLP [3] : Enhances VSLNet with clip features learned through egocentric video-language pretraining

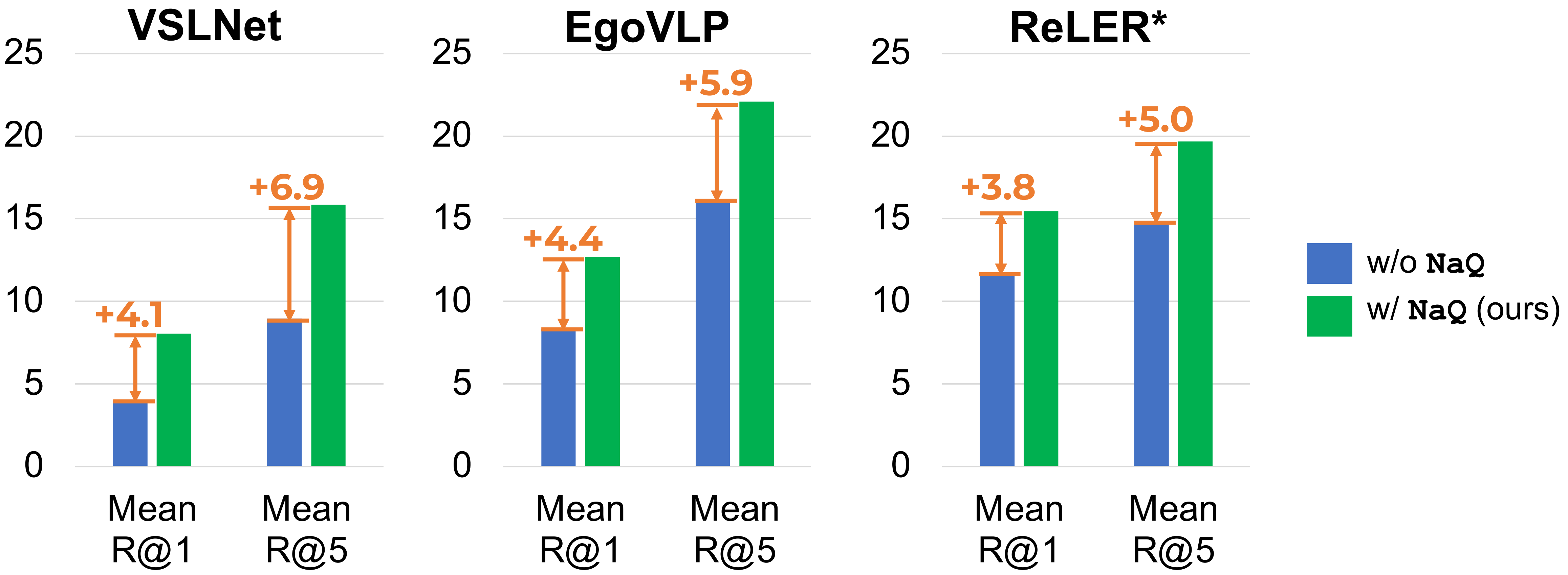
ReLER* [4] : Improves over VSLNet architecture + uses video-level data augmentation

*we further improve the ReLER baseline using EgoVLP features

[1] Grauman, Kristen, et al. "Ego4d: Around the world in 3,000 hours of egocentric video." *CVPR 2022*
[2] Zhang, Hao, et al. "Span-based Localizing Network for Natural Language Video Localization." *ACL 2020*
[3] Lin, Kevin Qinghong, et al. "Egocentric video-language pretraining." *NeurIPS 2022*
[4] Shao, Jiayi, Xiaohan Wang, and Yi Yang. "ReLER@ ZJU Submission to the Ego4D Moment Queries Challenge 2022." *arXiV 2022*

Experimental results

NaQ augmentation ***consistently*** and ***significantly*** enhances all baselines



Our approach improves NLQ performance by **up to 7%** absolute mean recall

*we further improve the ReLER baseline using EgoVLP features

Experimental results

NaQ sets the state-of-the-art results on the public Ego4D NLQ leaderboard

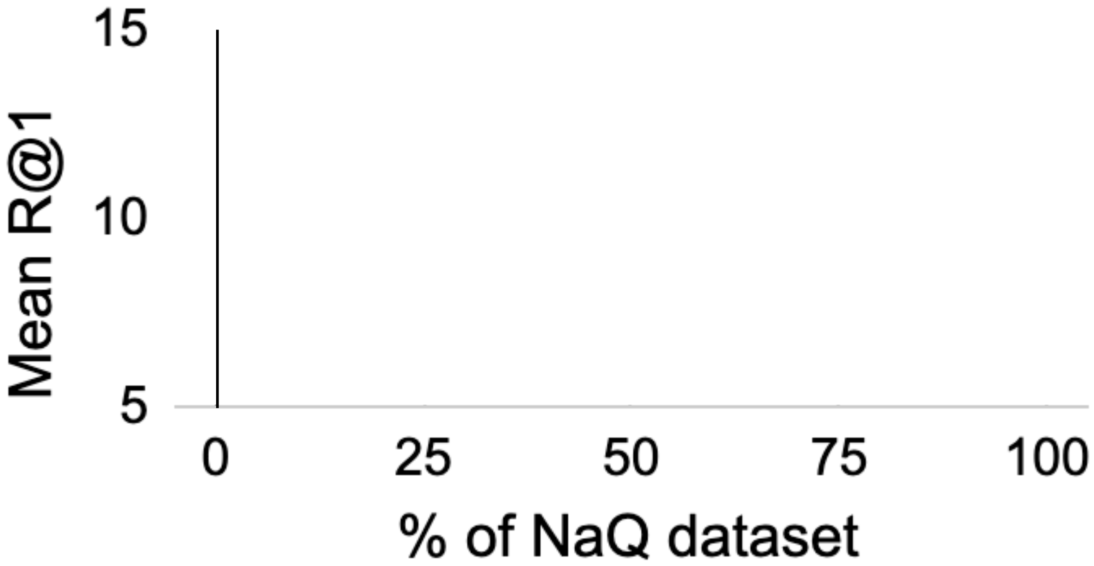
Method	R@1 IoU=0.3	R@1 IoU=0.5	Mean R@1†	R@5 IoU=0.3	R@5 IoU=0.5
`NaQ++` (ours)‡	**21.70**	**13.64**	**17.67**	25.12	16.33
`NaQ` (ours)	**18.46**	**10.74**	**14.59**	21.50	13.74
InternVideo [5]	16.46	10.06	13.26	22.95	16.11
Badgers@UW-Mad. [27]	15.71	9.57	12.64	**28.45**	**18.03**
CONE [18]	15.26	9.24	12.25	26.42	16.51
ReLER [24]	12.89	8.14	10.51	15.41	9.94
EgoVLP [23]	10.46	6.24	8.35	16.76	11.29
VSLNet [38]	5.42	2.75	4.08	8.79	5.07

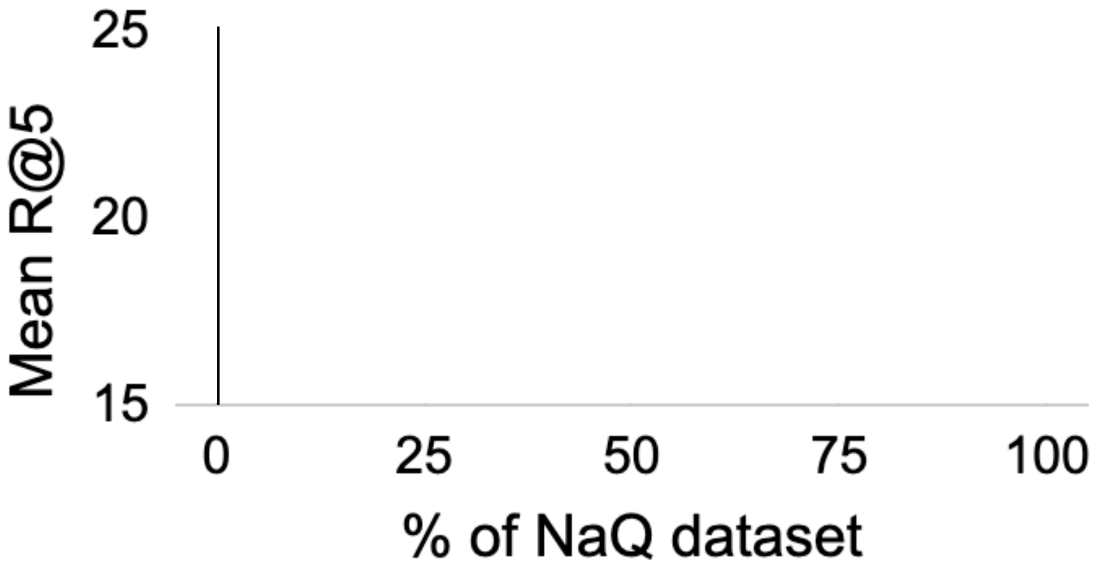
† Mean R@1 is the primary metric for deciding challenge winners
‡ NaQ++ combines winning entries from prior challenges and NaQ to achieve SoTA

Our approach improves NLQ SotA by 4.5% absolute mean recall @ 1

Experimental results

NaQ performance scales with the number of narrations used for training

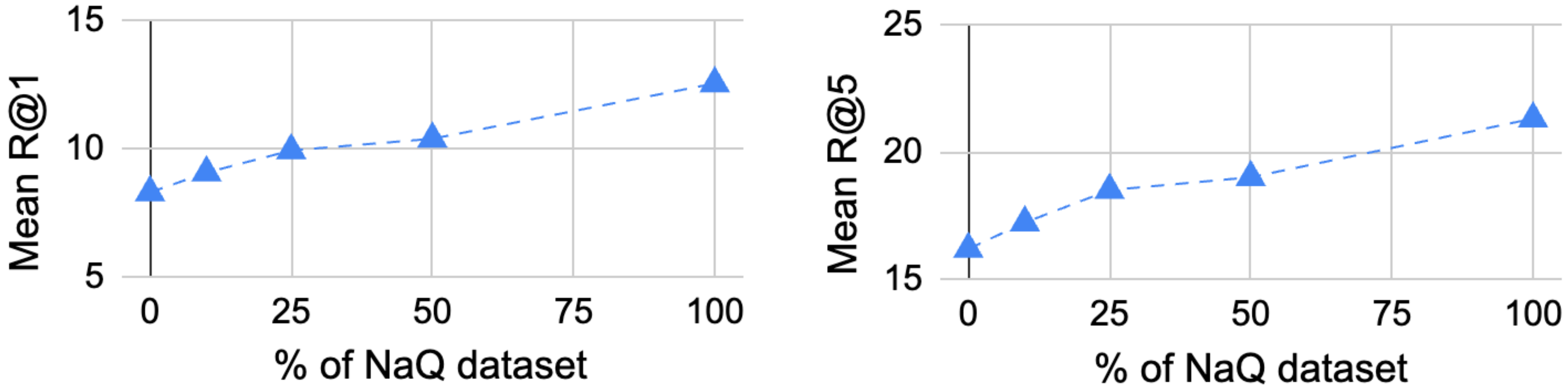




narrations used for NaQ augmentation

Experimental results

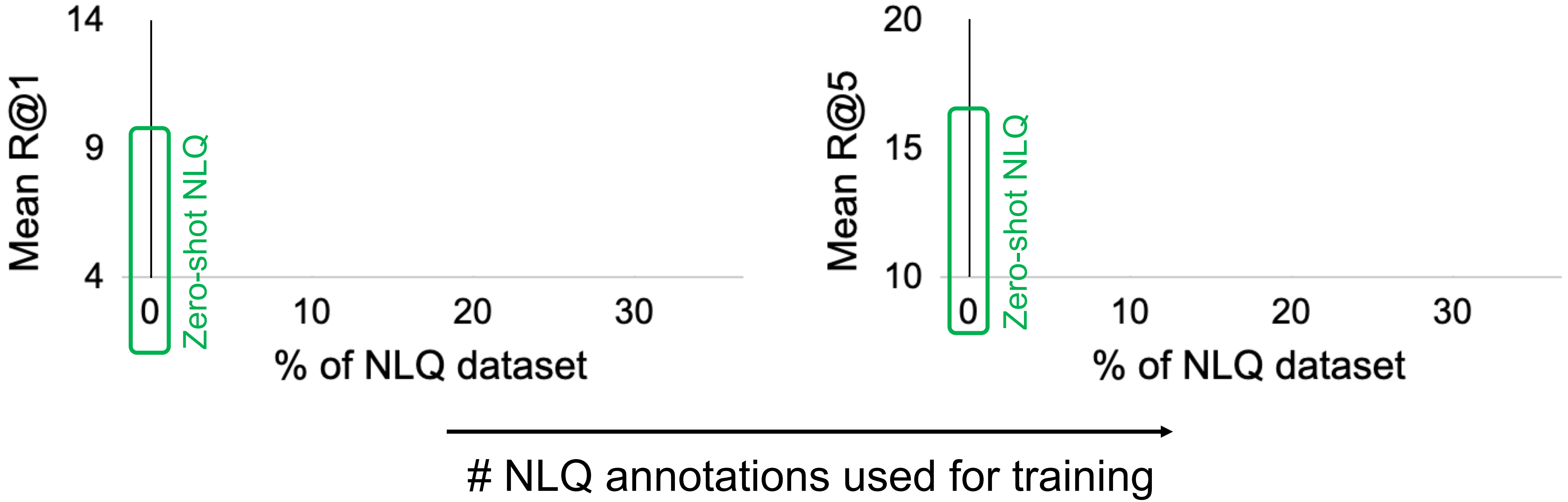
NaQ performance scales with the number of narrations used for training



narrations used for NaQ augmentation

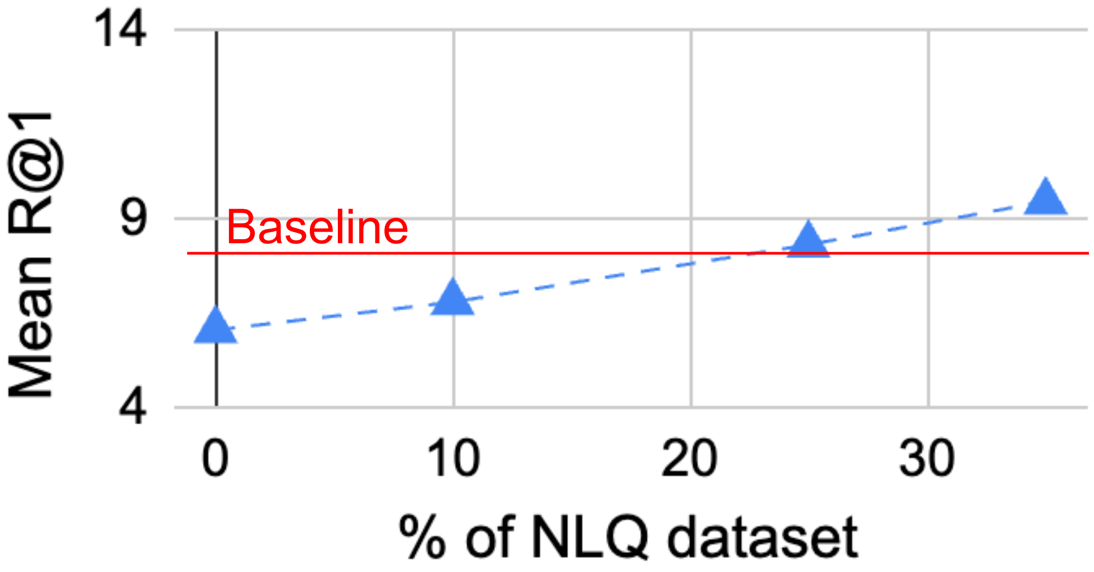
Experimental results

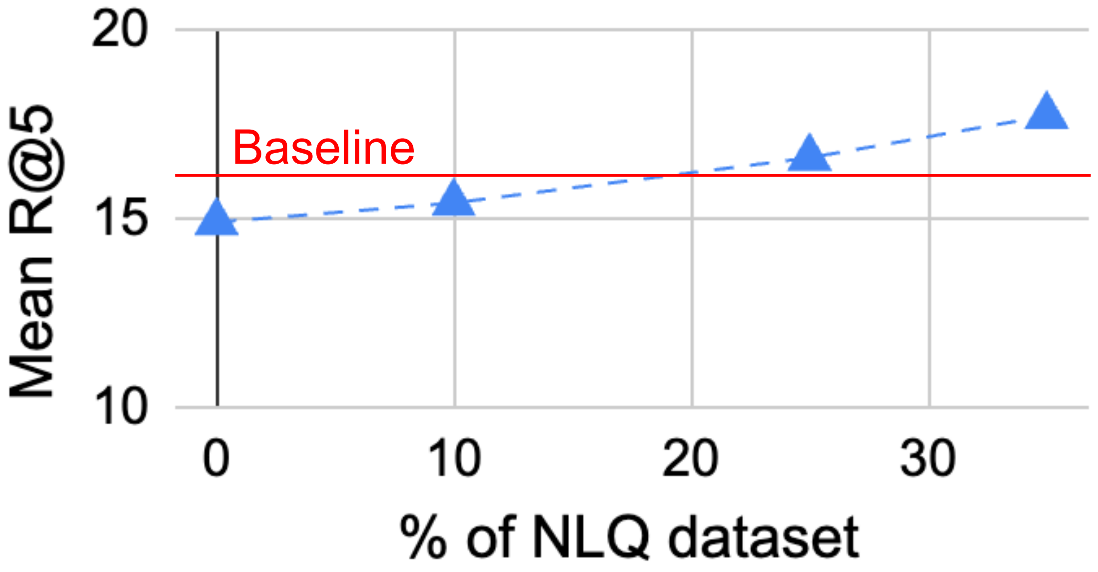
NaQ facilitates zero-/few-shot NLQ



Experimental results

NaQ facilitates zero-/few-shot NLQ

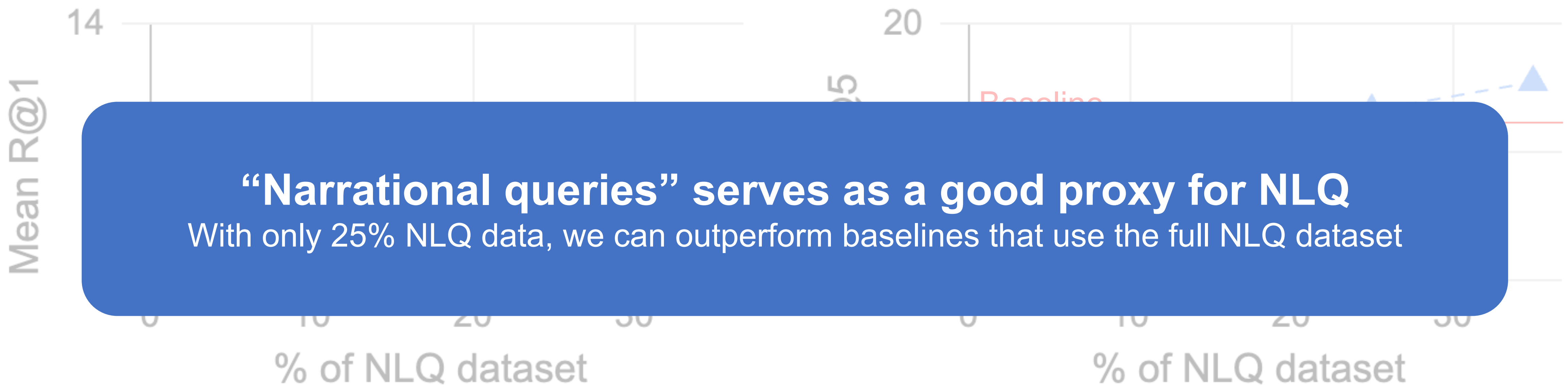




NLQ annotations used for training

Experimental results

NaQ facilitates zero-/few-shot NLQ

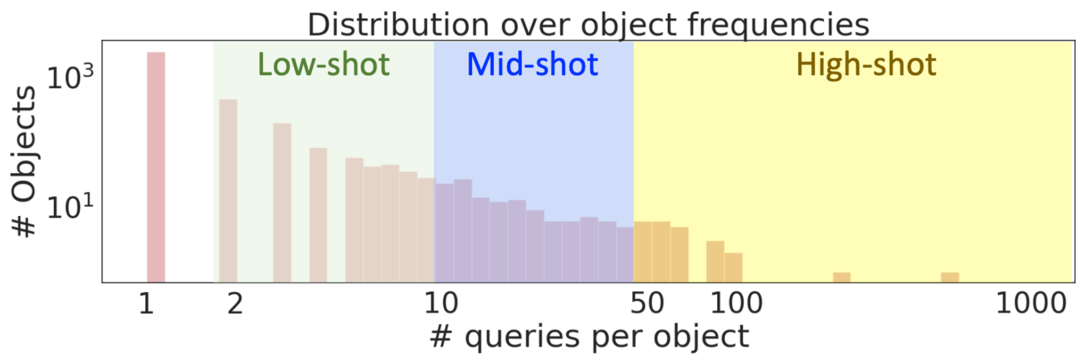


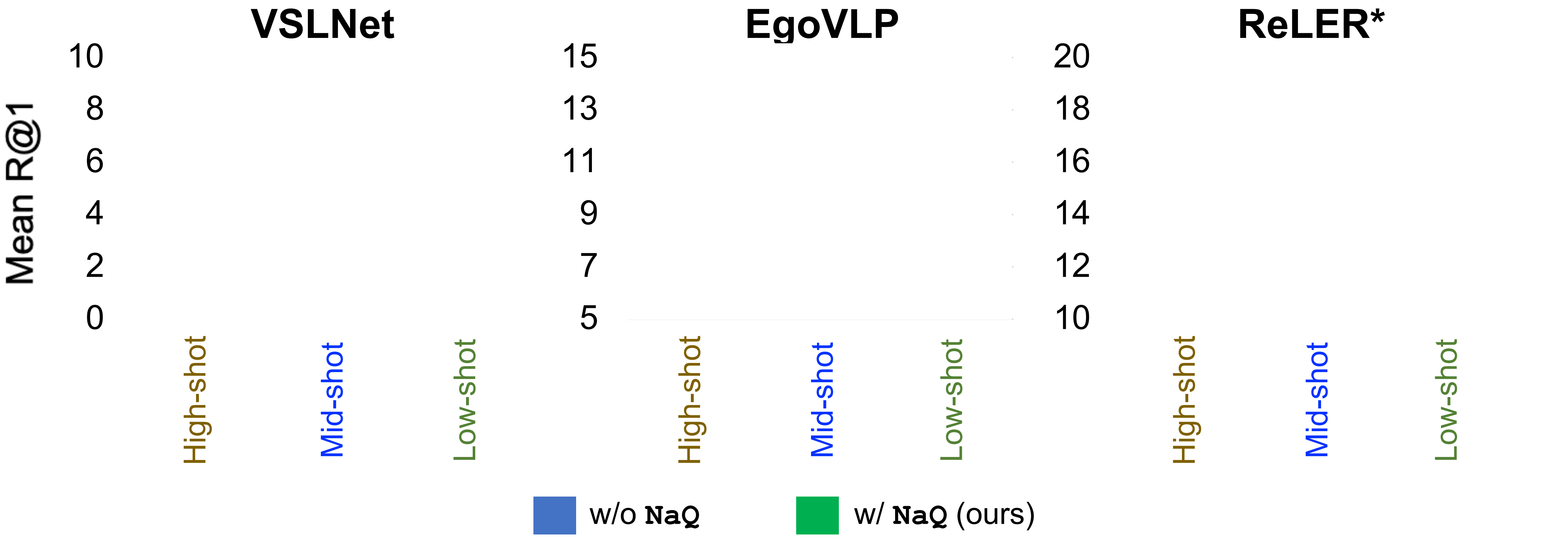
“Narrational queries” serves as a good proxy for NLQ

With only 25% NLQ data, we can outperform baselines that use the full NLQ dataset

NLQ annotations used for training

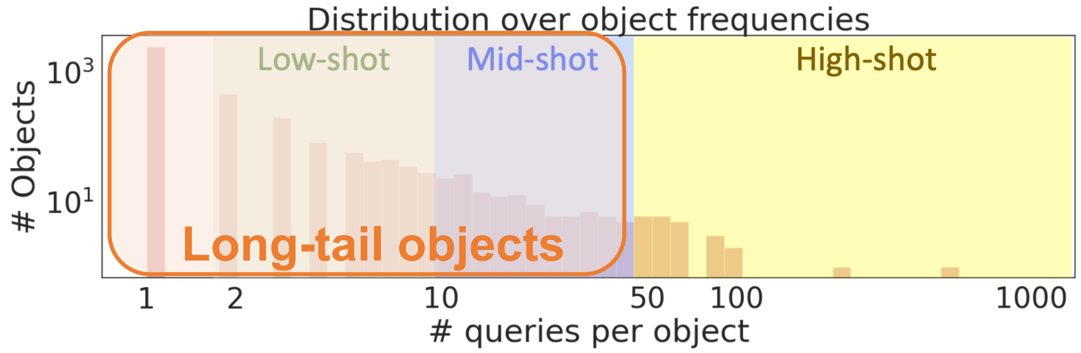
Experimental results



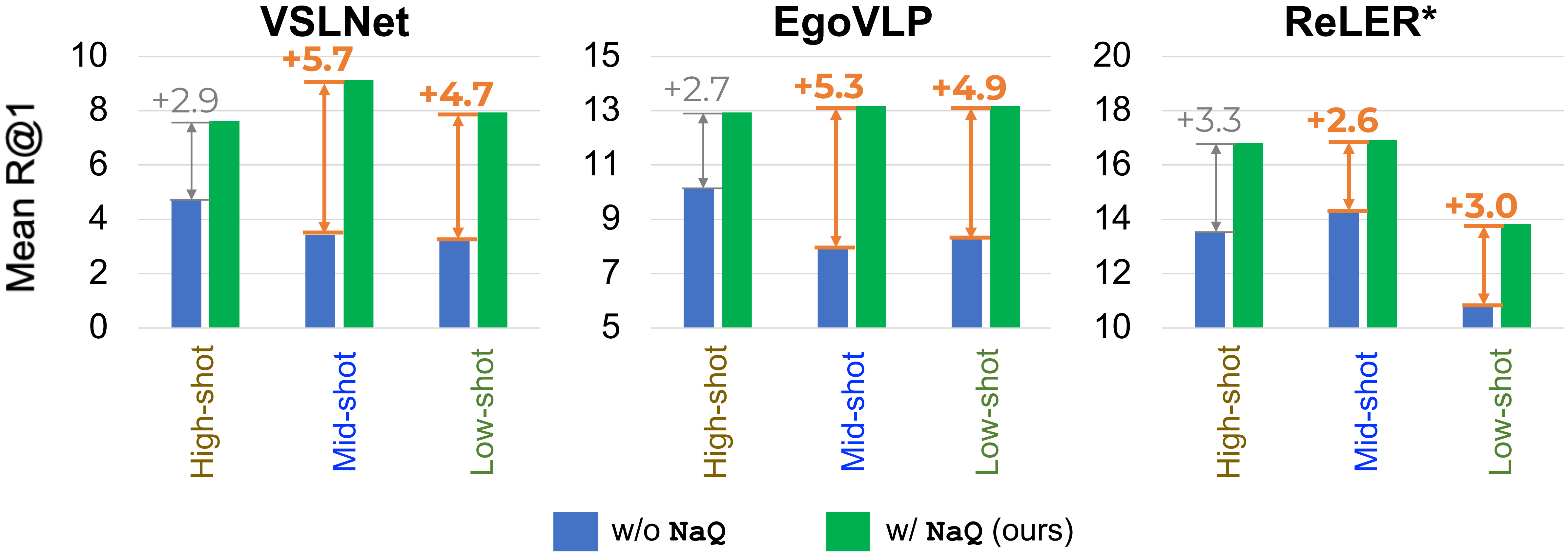


*we further improve the ReLER baseline using EgoVLP features

Experimental results

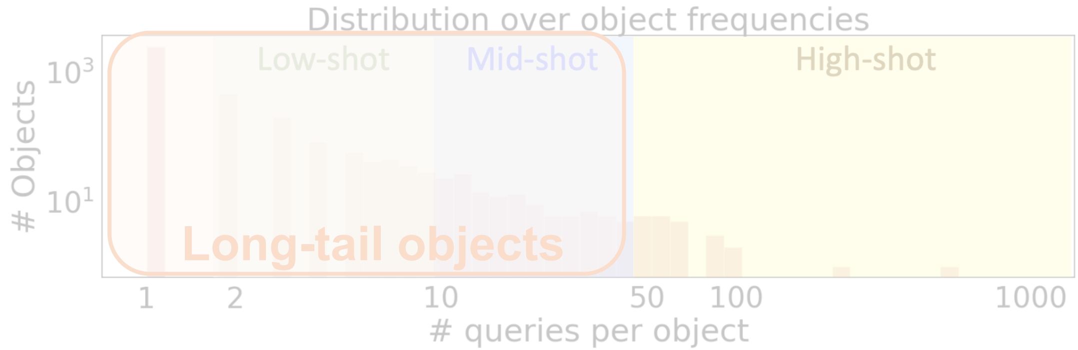


NaQ significantly improves responding to queries about **long-tail objects**

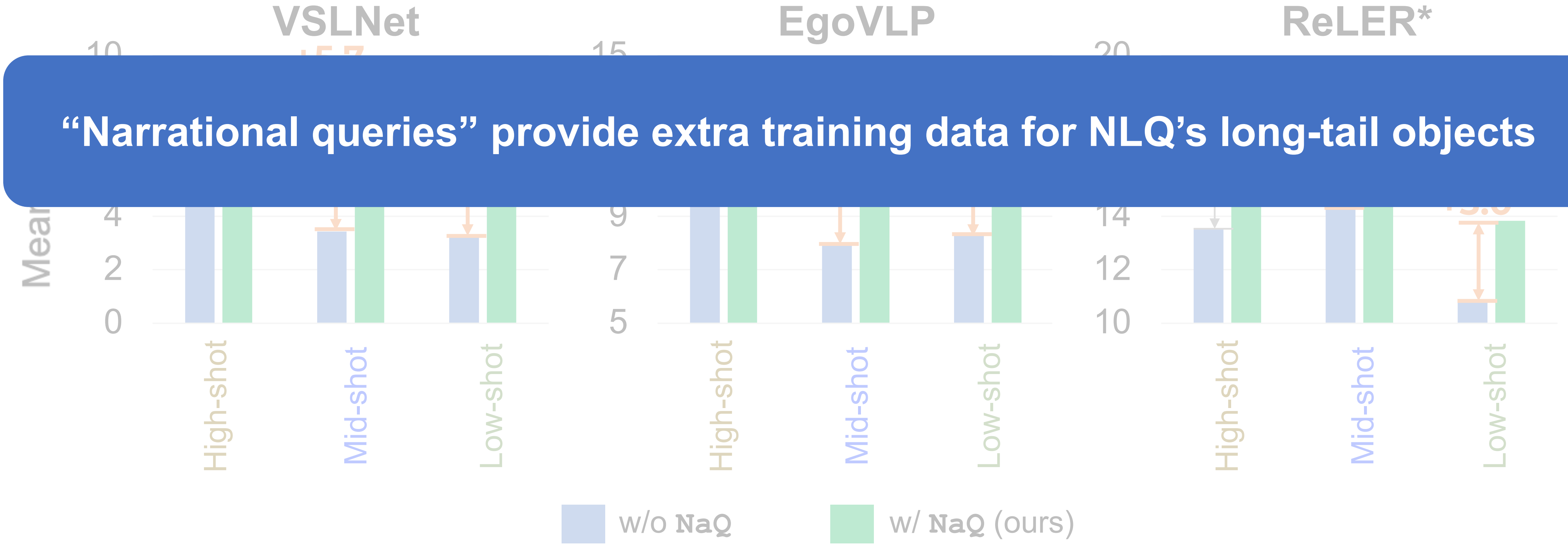


*we further improve the ReLER baseline using EgoVLP features

Experimental results



NaQ significantly improves responding to queries about **long-tail objects**



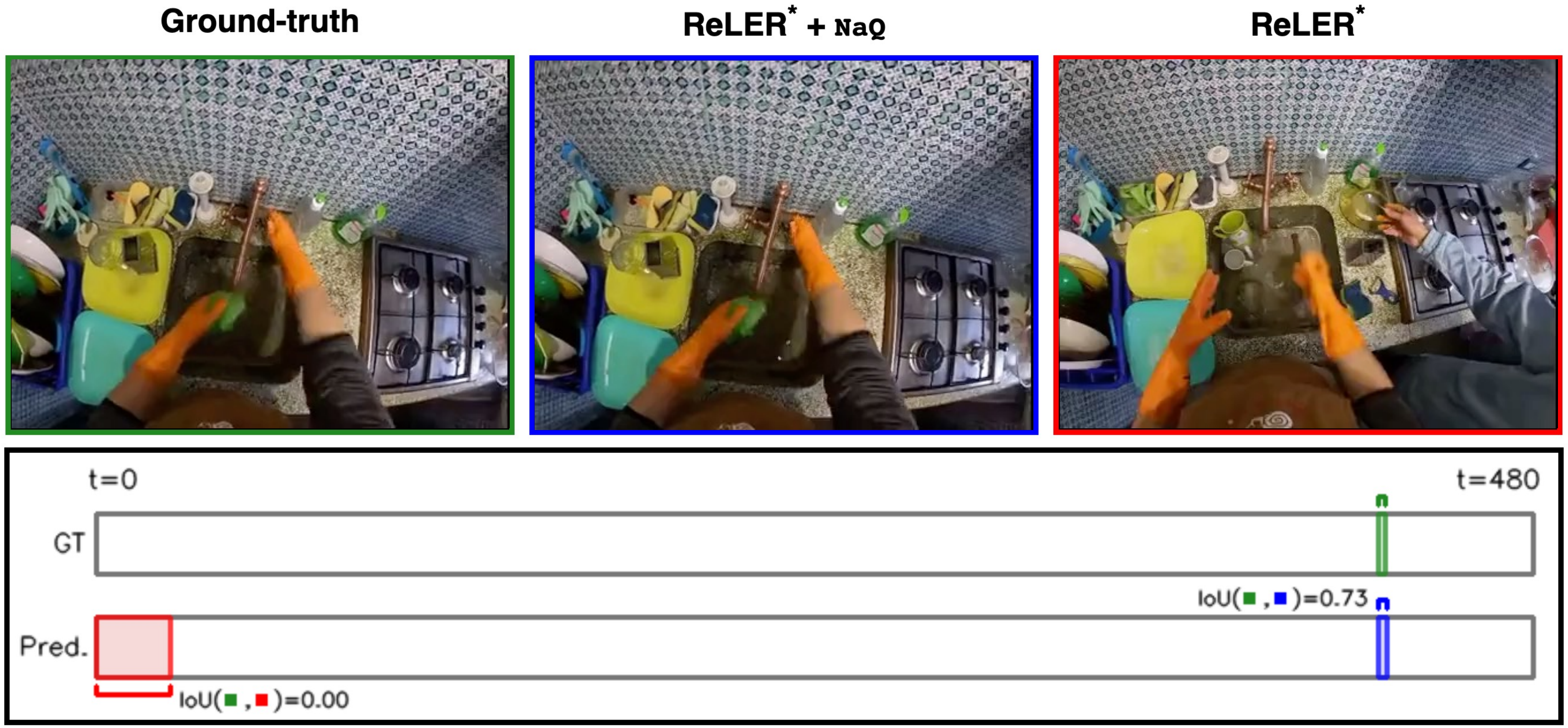
"Narrational queries" provide extra training data for NLQ's long-tail objects

*we further improve the ReLER baseline using EgoVLP features

Qualitative results

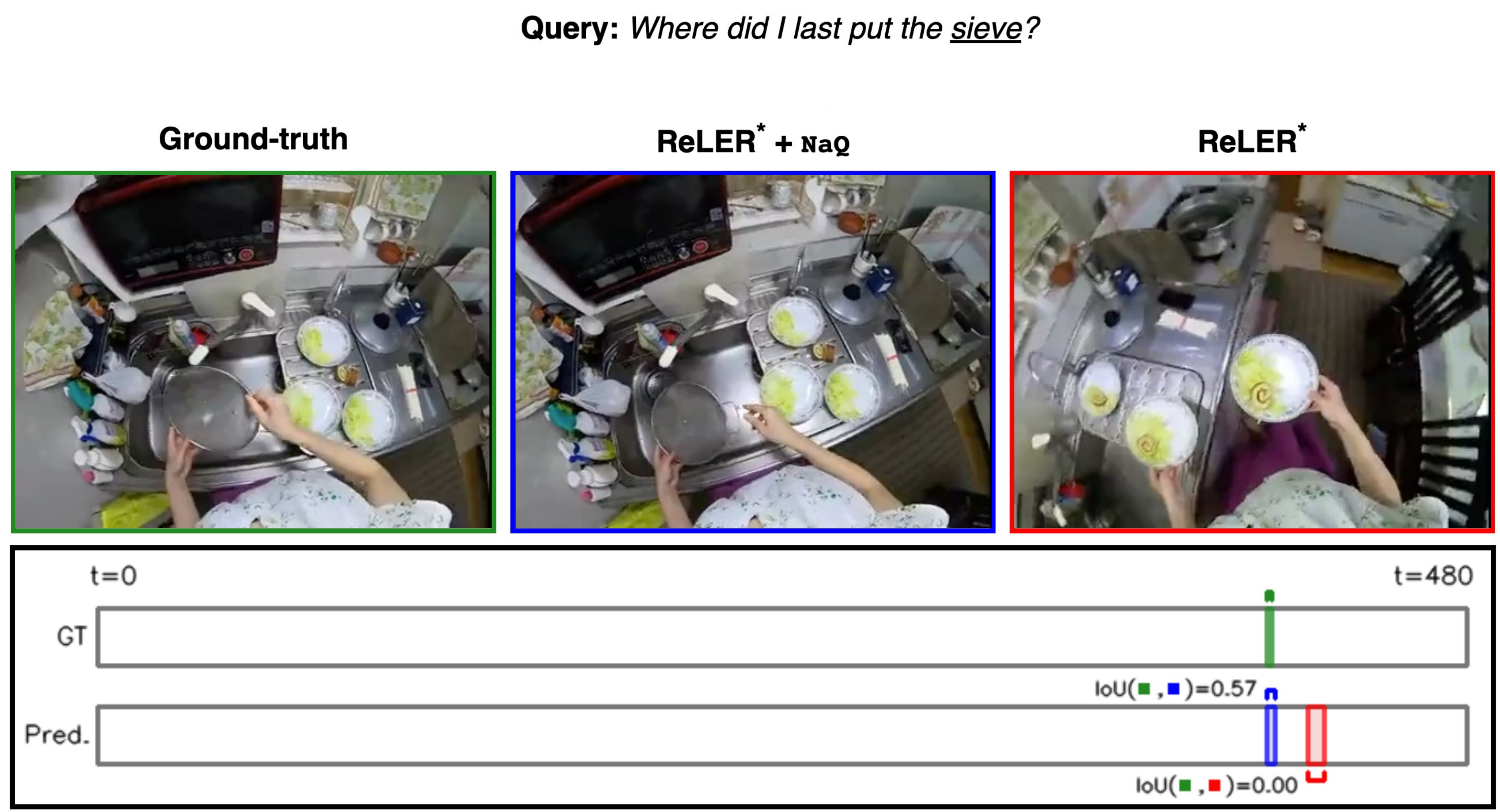
NaQ succeeds, while baseline fails, to reason about the long-tail object “soap”

Query: *Where was the soap before I picked it?*



Qualitative results

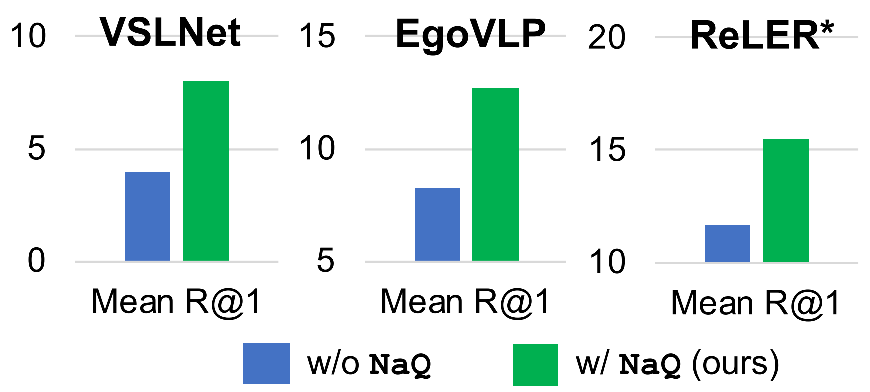
NaQ succeeds, while baseline fails, to reason about the long-tail object “sieve”



Conclusion

NaQ: Simple-yet-effective augmentation strategy for Episodic Memory NLQ

Improvements for multiple NLQ methods

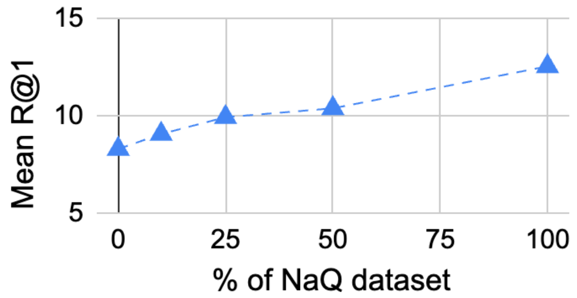


Obtains SotA on Ego4D NLQ

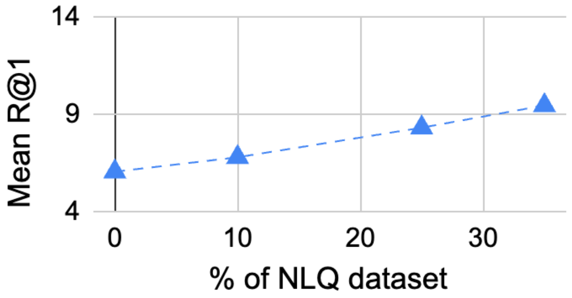
Method	R@1 IoU=0.3	R@1 IoU=0.5	Mean R@1†
NaQ++ (ours)‡	21.70	13.64	17.67
NaQ (ours)	**18.46**	**10.74**	**14.59**
InternVideo [5]	16.46	10.06	13.26
Badgers@UW-Mad. [27]	15.71	9.57	12.64
CONE [18]	15.26	9.24	12.25
VSLNet [38]	5.42	2.75	4.08

Interesting properties from using narrations-as-queries

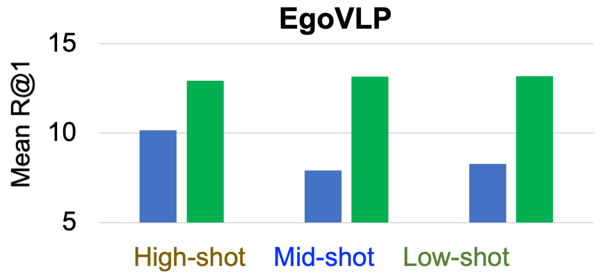
1) Data scaling



2) Zero-/few-shot NLQ



3) Benefits long-tail objects



JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

NaQ: Leveraging Narrations as Queries to Supervise Episodic Memory

Santhosh Kumar Ramakrishnan[1]   Ziad Al-Halah[2]   Kristen Grauman[1,3]

Poster session: TUE-PM-245

Project page: https://vision.cs.utexas.edu/projects/naq/

Code: https://github.com/srama2512/NaQ

1 TEXAS The University of Texas at Austin

2 THE UNIVERSITY OF UTAH

3 Meta AI