We thank all reviewers unanimously acclaim on both the novelty and performance of EventRR.

Dear Reviewer 1, #1.1 Revisit Refer-Aware Positional Encoding. **ReferPE is learnable and implemented as `nn.Embedding(50)`.** Besides, since our Referential Event Graph (lets denote it as `G`) is a **rooted DAG**, for each node, its shortest path connecting to the root node (whose index is 0 in `networkx` implementation) can always be found. We implement as `networkx.single_target_shortest_path_length(G, 0)`. Finally, **each node is associated with a path length integer ($\geq$ 0, <max depth), we then add the corresponding learnable embedding to the node feature.** #1.2 Unfair multi-task v.s. task-specific training, but we validate our event-centric motivation. **UniRef [ICCV'23] and GLEE [CVPR'24] are trained on unified multi-task dataset**, including Referring Detect/Segment (like large-scale Visual Genome), Uni-modal Detect/Segment, **while EventRR is trained only on RIOS/RVOS**. More referring(uni-modal) samples surely improve cross-modal alignment(segment accuracy). **Although data is unfair, in Figure 1, we empirically found UniRef/GLEE fails on the "eventful" samples while EventRR succeeds.** For example, the first "sheep" example has event-attribute information("moving towards right"). Since GLEE/UniRef uses online-RVOS-inference, for first few frames, they don't know which sheep is later moving right. The second "fish" example has event-event relation information ("left" then "right"), GLEE/UniRef also cannot capture this temporal context. They also fail the other two "eventually"/"when" samples. In Figure 2, interpretable visualizations of the reasoning process also validate our event-centric design.

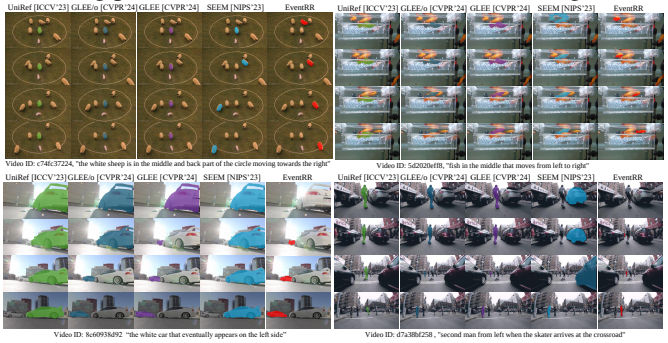

Figure 1. Visualizations of UniRef[ICCV'23](UniRef++-R50),SEEM[NIPS'23](SEEM-v0-FocalT),GLEE[CVPR'24](GLEE-Pro-Stage2-EVA02L), EventRR(VSwinB+RIOS-Pretrain) on four "eventful" samples in Ref-Youtube-VOS. GLEE/o means online-inference[1], while GLEE means per-frame inference. EventRR succeeds on these examples, which validates our event-centric design. We will cite SEEM/GLEE/UniRef/SLiMe/MLRL.

Dear Reviewer 2, #2.1 Capture-then-Align hypothesis. Aligning cross-modal event information(event attribute / event-event relation) is necessary for accurate RVOS. But model must first capture that information in each modal, which requires visual temporal context aggregation. **By no means can per-frame methods aggregate along temporal axis, let alone later alignment.** To validate, we per-framely evaluate two SOTA image methods: SEEM [NIPS'23] and GLEE [CVPR'24](Reviewer1) on Ref-Youtube-VOS. As shown in Figure 1, both SEEM and GLEE show: **(a) event-unawareness.** Their knowledge is limited to information like "middle, white, sheep, car", which biases them to appearance/spatiality. Their predictions on "fish" example are "middle" fish, but not the "moving from left to right" "middle" fish. **(b) lack of temporal consistency.** In "sheep"/"man" example, SEEM/GLEE predicts different instance across different frame. (Due to space limitation, we will carefully correct these typos and polish Figure 4(add more illustrations for each symbol).).

Figure 2. Interpretable visualization of the reasoning process for the "sheep" sample. (left) Referential Event Graph (REG) of the text. (right) Cross-attention-heatmap of the temporal query with maximum score when reasoning (TCRR) arrives at "right-04" and "move-01", respectively. From focusing on "right" region to "moving towards right" region, **EventRR successfully aligns the event attribute information in video and that in text**.

Dear Reviewer 3,

#3.1 Cursory, unspecific, and subjective evaluations.

For 1)(subjective, cursory, unspecific), we **very clearly** stated our motivations and corresponding efforts. Please scrutinize line032-093, line108-121, 141-163, line244-313.

For 2)(unspecific), EventRR achieves **much higher J&F** on Ref-Youtube-VOS than MLRL-CVPR22 (**59.2/66.9** > **49.7**). Besides, MLRL is **not open-sourced**. Finally, EventRR is based on the **modern DETR regime**, which apparently necessities different fusion strategies from old MLRL framework. Please scrutinize line206-242, line266-294.

For 3)(cursory, unspecific), we **very clearly** detailed our model and implementation in line330-350, 380-390.

#3.2 Impractical and perfunctory NLP-request. For 4), BERT/RoBERTa/AMRBART all use BPE tokenization(*Tokenizer*) and learnable embeddings. Only difference is their vocabulary size, i.e. `nn.Embedding(vocab_size)`. Besides free-form English words, **AMRBART also knows how to encode AMR concepts//roles** like:) walk-01 | walk-02 // :destination | :ARG0 | :time. Applying BERT to EventRR is **completely impractical and cannot be coded.**

#3.3 (VERY IMPORTANT) We DID-NOT claim dataset contribution in our paper and DID indicate in the form.

[1]In GLEE[CVPR'24] paper, during RVOS-online-inference, for each current frame query, its "similarity (not specified in paper)" to the previous frame selected query is added to its referring score. However, by the rebuttal ddl, GLEE hadn't released the online-inference code, so we took the query feature before the decoder norm and implemented the similarity measure as `nn.functional.cosine_similarity`.