

Peer Review

Review of: "LLaVA-MR: Large Language-and-Vision Assistant for Video Moment Retrieval"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper takes Mr. BLIP, a generative BLIP-2-based model that treats video moment retrieval as sequence-to-sequence generation, and argues that its weak point is sparse frame sampling forced by the LLM context. LLaVA-MR adds three pieces on top. Dense Frame and Time Encoding (DFTE, §3.2) samples more frames and interleaves each frame with a time token, choosing relative frame indices when the sampling rate is at least one frame per second and rounded timestamps when it is lower. Informative Frame Selection (IFS, §3.3) ranks frames by a Gaussian-smoothed L2 distance between adjacent image-encoder features and keeps the top k as key frames. Dynamic Token Compression (DTC, §3.4) then shrinks the Q-Former output of the non-key frames, either by average pooling or by keeping the half of the 32 queries with the highest variance across frames. Training uses LoRA (§3.5, §4.1, "0.63% of the total parameters").

The headline claim is state of the art on Charades-STA and QVHighlights (Table 1), plus ActivityNet Captions in the appendix (Table 6). On the QVHighlights test split, the abstract reports gains of 1.82 in R1@0.5 and 1.29 in mAP@0.5. A four-step ablation on Charades-STA (Table 2) attributes the gain from 65.48 to 70.65 R1@0.5 to denser frames, time encoding, and IFS&DTC, and Tables 3, 4, 5, 7, and 8 study the individual design choices.

What I value here is concrete. The ablation chain is internally consistent, which is rarer than it should be: Table 2 row (b) (66.37 / 44.65 / 57.50) equals the "Timestamps" row of Table 3 for Charades-STA, the "Relative frame indices" row of Table 3 (67.83) equals "No special tokens" in Table 4, "B/E for both" in Table 4 (68.52) equals Table 2 row (c) and "No Compression" in Table 5, and the 1:2 row of Table 7 (70.65 /

49.58 / 59.78) equals Table 2 row (d). I could trace the whole path from baseline to final model through the tables. The QVHighlights test numbers come from the hidden-label evaluation server (§4.1, footnote 1), so they cannot have been tuned on. The variance-based DTC idea is simple, cheap, and well motivated, and the paper is honest about failure cases (Figure 4, example 5; Figure 7, examples 7 and 8).

2. Whether the conclusion holds on the data

The main claim, that LLaVA-MR improves over Mr. BLIP on the standard benchmarks, is supported in direction on every split the paper reports. I checked the differences against Mr. BLIP from Table 1: QVHighlights validation R1@0.5, $78.13 - 76.13 = 2.00$; QVHighlights test R1@0.5, $76.59 - 74.77 = 1.82$; R1@0.7, $61.48 - 60.51 = 0.97$; mAP@0.5, $69.41 - 68.12 = 1.29$; mAP@0.75, $54.40 - 53.38 = 1.02$. Charades-STA R1@0.5 gives $70.65 - 69.31 = 1.34$, and ActivityNet Captions (Table 6) gives $55.16 - 53.92 = 1.24$. The gains quoted in §4.2 and Appendix 6.1 match these differences. Consistent direction on three datasets and on a server-scored test split is real evidence.

The size of the gains is where I would ask for more care.

Units. The abstract and §4.2 write "an improvement of 1.82%". These are absolute differences in percentage points. As a relative change, $1.82 / 74.77 = 2.4\%$. I would write "+1.82 points" throughout.

Run-to-run variance. Every number in the paper is a single run with no standard deviation and no seed. To give a sense of scale, I computed a binomial standard error for R1@0.5 from the split sizes in §4.1. Charades-STA test, 3,720 annotations: $\sqrt{(0.7065 \times 0.2935 / 3720)} = \sqrt{(5.57 \times 10^{-5})} \approx 0.75$ points. QVHighlights test, 1,542 queries: $\sqrt{(0.7659 \times 0.2341 / 1542)} = \sqrt{(1.16 \times 10^{-4})} \approx 1.08$ points. The Charades-STA margin over Mr. BLIP (1.34) and the margin over InternVideo2-6B ($70.65 - 70.03 = 0.62$) are therefore within about two standard errors of sampling noise alone, before any training variance. A paired test is more sensitive than this rough bound, and that is exactly why I would like to see one: per-query predictions of LLaVA-MR and Mr. BLIP are both available to the author, so a paired bootstrap or McNemar test on R1@0.5 would settle the question cheaply. Three seeds for the final model would do the same for training variance.

Post-processing. §4.1 and Appendix 7 describe heuristics that repair malformed LLM outputs. The note under Table 8 says those rows were obtained without post-processing. The best Table 8 row ($N = 80$, $k = 32$) gives 76.97 R1@0.5 on QVHighlights validation, while Table 1 reports 78.13 for the same split. If that row is the final configuration, then $78.13 - 76.97 = 1.16$ points comes from post-processing alone, which is comparable to the claimed margin over Mr. BLIP. I could not tell from the text whether the Mr. BLIP

numbers in Table 1 were taken from its paper or re-run with the same post-processing. This needs one sentence and, ideally, one extra row.

Validation-set selection. N and k for QVHighlights were chosen on the validation set (Table 8), and the validation result is then reported in Table 1 next to baselines. The test split is the clean comparison; the 2.00-point validation gain is somewhat optimistic. Charades-STA has no separate validation split in §4.1, so I would state on which data the Charades hyperparameters (Figure 6, Table 7) were chosen.

A contradiction in §4.3. The text says that for QVHighlights "using frame indices as time tokens led to better performance, improving the R1@0.5 score by +1.29". Table 3, QVHighlights block, shows the opposite: Timestamps 76.06 against Relative frame indices 74.77, a difference of $76.06 - 74.77 = 1.29$ in favour of timestamps. The number is right and the label is wrong. Since the whole sampling-rate rule of §3.2 rests on this sentence, it should be corrected.

The sampling-rate rule itself. Table 3 supports "indices on Charades-STA, timestamps on QVHighlights", but dataset and sampling rate change together, so the table cannot isolate the rate. A controlled test would vary N on one dataset (say 20, 30, 60 frames on Charades-STA, crossing the 1 frame/s threshold) and compare both encodings at each N.

"Greater advantage on longer videos" (§1, third contribution; §4.2). The test gains over Mr. BLIP are 1.34 on Charades-STA (about 30 s videos), 1.82 on QVHighlights (150 s), and 1.24 on ActivityNet Captions (about 2 minutes, Table 6). That is not monotone in length, and the differences between the gains are smaller than the noise estimated above. I would soften this to an observation on QVHighlights.

Ablation granularity. Table 2 step (c) bundles relative indices and special tokens, and step (d) bundles IFS and DTC. Table 4 separates the first pair. For the second, I could not find IFS without DTC, or DTC applied to all frames without IFS. Table 7 uses the variance-based method only. Given that IFS&DTC is the paper's main contribution (+2.13 in Table 2: $70.65 - 68.52$), separating the two would make the claim much stronger.

Smaller points. The inference-time claim (1.31 s against 1.57 s, §4.2) is an average of 10 runs with no spread and no stated frame count. Table 1 shows mAP for Charades-STA (69.96 and 39.66) and mIoU for QVHighlights validation (71.85) only for LLaVA-MR, so those cells are not comparisons. The qualitative examples on news and film in Appendix 8 are informative, but they are selected examples and do not support a generalisation claim on their own. In Figure 4, example 4, the extra unannotated segment would count as a false positive under the reported metrics; it is worth saying so.

3. Reproducibility

Some of what is needed is there: optimizer and schedule (AdamW, learning rate warmed from $1e-8$ to $3e-4$ over 10% of iterations, cosine decay), frames per video (60 for Charades-STA, 80 for QVHighlights), epochs (up to 50), batch sizes (16 and 32), gradient accumulation (4), hardware and GPU hours ($4 \times$ A100-80GB, about 70 GPU hours; $8 \times$ A100-80GB, about 33 GPU hours), and the chosen N, k and T (§4.3, Appendix 6.2: N = 60, k = 32 on Charades-STA, N = 80, k = 32 on QVHighlights, T = 16). Dataset splits are standard and described.

The following I could not find:

- The exact BLIP-2 checkpoint, including which LLM it uses and which image encoder. "BLIP-2" covers several variants with very different language models.
- LoRA rank, alpha, dropout, and target modules. Only the 0.63% figure is given.
- Beam size, maximum generation length, and any length penalty for beam-search decoding (§3.5).
- Gaussian filter width for IFS (§3.3), and whether the DTC variance is computed per video over all N frames (as Figure 3 says) or over non-key frames only.
- The text prompt given to the LLM, and the exact post-processing code. Appendix 7 describes it in prose and says it normalised "over 50,000 samples", but the rate of malformed outputs before repair is not reported.
- Random seeds and the number of training runs.
- Code and weights: the abstract says "will be open-sourced upon acceptance", so at present nothing is released.

One detail looked odd to me. Charades-STA training (60 frames, 12,408 annotations) took about 70 GPU hours, while QVHighlights (80 frames, 7,218 training queries, batch 32) took about 33. That may be right, but a word on why would help anyone trying to budget a reproduction.

4. What specifically to fix

1. **§4.3, Table 3 discussion.** Replace "using frame indices as time tokens led to better performance" for QVHighlights with "using timestamps"; Table 3 shows 76.06 vs 74.77.
2. **Abstract and §4.2:** report gains as points, not percent (1.82 points absolute is 2.4% relative).
3. Abstract says LLaVA-MR outperforms "11 state-of-the-art methods"; §5 says "ten leading methods". Table 1 lists 11 distinct methods by my count (Moment-DETR, EaTR, QD-DETR, UnLoc-L, CG-DETR,

Mr. BLIP, SeViLa, UniVTG, UniMD+Sync, InternVideo2-1B, InternVideo2-6B). One number, please.

4. **Table 1.** I would add a paired significance test against Mr. BLIP on each test split, or mean $\pm$ std over three seeds. Either one.
5. State whether Mr. BLIP in Table 1 used the same post-processing, and add a Table 1 row for LLaVA-MR without post-processing so the 76.97 vs 78.13 gap (Table 8 vs Table 1, QVHighlights validation, R1@0.5) is explained.
6. **Table 2 / Table 5:** add IFS alone and DTC alone, so step (d) is not a bundle.
7. For the §3.2 rule, a controlled sweep of N on Charades-STA with both time encodings would turn a plausible story into a tested one.
8. **§4.3, IFS design paragraph:** "As illustrated in Figure 4" should be Figure 6, which is the hyperparameter figure; Figure 4 is the qualitative one.
9. **References:** entries [3] and [9] are the same Lei et al. QVHighlights paper (arXiv and NeurIPS versions), so Appendix 6.2 cites "QVHighlights[3]" while §4.1 cites [9]. One entry is enough. §4.1 also says "QVHighlights Captions", which I think is a mix-up with ActivityNet Captions.
10. Give the inference-time comparison (§4.2) a frame count, a standard deviation, and the decoding settings used.
11. **§3.5 and §4.1:** list the BLIP-2 checkpoint, LoRA hyperparameters, beam size, Gaussian filter width, and the prompt template. A small table is enough.
12. Report the share of raw outputs that needed repair in Appendix 7.
13. Soften "demonstrating a greater advantage on more complex, longer videos" (§1) unless a length-stratified analysis is added; the gains of 1.34, 1.82, and 1.24 do not order by video length.
14. A short note on naming. The model is built on BLIP-2, not LLaVA, and readers may expect a LLaVA backbone from the title. A sentence in §3.1 would prevent that.

5. Recommendation

Minor revisions. The method is sensible, the ablation tables reconcile with each other, and the improvement over the strongest baseline is consistent across three datasets, including a server-scored test split. The fixes are mostly textual (one reversed sentence in §4.3, units, a method count, a wrong figure reference, a duplicated citation) plus a significance test, a post-processing control, and the missing configuration details. None of them requires new architecture work, and all of them would make a solid paper easier to trust and to reproduce.

Declarations

Potential competing interests: No potential competing interests to declare.