

Peer Review

Review of: "CM3T: Framework for Efficient Multimodal Learning for Inhomogeneous Interaction Datasets"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper proposes CM3T, a set of plug-in modules that adapt a frozen video transformer to new datasets and new input modalities without touching the backbone weights. Two modules are new. The **multi-head vision adapter** (MHVA, §3.2.2) splits the input along the window dimension of Video Swin (or the patch dimension of ViViT) and gives each group of windows its own small parallel adapter, with a learnable scaling factor in place of a fixed one. The **cross-attention adapter** (CAA, §3.3) replaces the two linear layers of an adapter with a cross-attention block, taking the query and value from the MHVA bottleneck state and the key from the embedding of the extra modality; several CAAs are merged through AdapterFusion. A modified prefix tuning (ReLU instead of tanh, lower dropout) completes the set. The central motivation, drawn in Figure 1, is that existing adapter methods work well only on self-supervised or CLIP-pretrained backbones, and the author wants something that also works on ordinary supervised Kinetics-400 checkpoints.

The evidence comes from three datasets with very different settings: Epic-Kitchens-100 (egocentric actions, RGB plus optical flow plus audio), MPIIGroupInteraction (body language, RGB plus audio), and UDIVA v0.5 (personality regression, RGB plus transcript plus audio). The headline is that CM3T on a frozen Video Swin-B reaches 48.2% top-1 on EK-100 with 22.3% of the backbone's parameters trained (Table 4, row "CM3T", Multimodal input (EK-100)), against 49.6% for the much larger M&M Mix ensemble (Table 1), and that with a CLIP-B backbone it reaches 50.1% (Table 1), 0.65 mean MSE on UDIVA (Table 2), and 0.918 mAP on MPIIGI (Table 3).

2. Whether the conclusion holds on the data

What I value most here is the comparison in Table 1 that holds the pretraining fixed. When Dual-Path Adapters and ST-Adapters are run on a supervised ViT-B, they reach 35.8% and 34.3% (Table 1, rows "Dual-Path Adapters" and "ST-Adapters", backbone ViT-B), and the AdaptFormer-style adapter with prefix tuning on Video Swin-B reaches 28.7%, while MHVA + PT on the same Video Swin-B reaches 39.8%. That 11.1-point gap ($39.8 - 28.7$) between two adapter designs on an identical frozen backbone is the cleanest piece of evidence in the paper, and it directly supports the first contribution. The ViViT-B experiment in Table 6 adds a second backbone, which I appreciate; it shows the window/patch split is not tied to Swin.

I checked the arithmetic of the in-text comparisons against the tables, and most of it holds. The "only 1.4% worse" claim in §4.2.1 is $49.6 - 48.2 = 1.4$ (Table 1). The "0.7% lower than the fully finetuned model" in §4.2.3 is $48.9 - 48.2$ (Table 4, Multimodal input (EK-100)). The 8.4-point modality gain in §4.2.4 is $48.2 - 39.8$ (Table 5), and the 2.9 points for M&M Mix over MTV-B is $49.6 - 46.7$. The "more than 13 times less" trainable parameters also checks out roughly: 22.3% of 88M is about 19.6M, and $>300\%$ of 88M (M&M Mix, Table 5) is $>264\text{M}$, a ratio above 13.

Where the conclusions run ahead of the data is mostly a matter of margins. Every number in Tables 1 to 6 is a single value with no standard deviation, no number of runs, and no seed. Several of the claimed wins are small. On MPIIGI, "better results" than FAt transformers (§4.2.2) rest on 0.901 against 0.899 mAP (Table 3), a difference of 0.002. On UDIVA, the Video Swin-B result is 0.69 against FAt's 0.72 (Table 2). The MMCA comparison in §4.2.4 is 48.2% against 47.7% (Table 5). And "even better performance than full-finetuning" for ViViT-B (§4.4) is 38.1% against 37.4% (Table 6). On EK-100's validation set, a difference of half a point can easily sit within run-to-run noise. I would not ask for the claims to be withdrawn. I would ask for them to be phrased as "on par" until repeated runs show the spread.

The parameter percentages do not reconcile with each other, and this matters because efficiency is the paper's main selling point.

- The abstract and §4.2.3 say CM3T uses 22.3% trainable parameters, and §4.2.3 says "again with only 22.3%" for UDIVA and MPIIGI. Table 4 lists **28.9%** for CM3T on both UDIVA and MPIIGI (rows "CM3T", Multimodal input (UDIVA) and (MPIIGI), column Trained params).
- MPIIGI uses one extra modality (audio, §4.1), EK-100 uses two (flow and audio). Yet the CAA overhead is larger for MPIIGI: $28.9 - 12.8 = 16.1\%$ against $22.3 - 12.8 = 9.5\%$ for EK-100. The full-finetuning rows

show the same pattern, 116.1% against 109.5%.

- §4.4.1 says each new modality needs 5.8M parameters. With the 88M reference from the Table 4 caption, that is 6.6% per modality, so two modalities would be about 13.2%, not the 9.5% implied for EK-100. Perhaps AdapterFusion or shared layers explain it, but I could not reconstruct the numbers from the text.

I also could not find parameter counts for the CLIP-B variants of CM3T, whose results (50.1%, 0.65, 0.918) carry the SOTA claims, nor for FAt transformers, so the "around 5 times less trainable parameters" statement in §4.2.2 cannot be checked from the paper. The trainable-parameter figures also seem to exclude the encoders that produce the audio, flow, and text embeddings (§4.4.1 says "given an embedding corresponding to features of the new modality"). That is a fair convention, but it should be stated, since M&M Mix processes optical flow with its own ViViT branch.

On compute, §4.4.1 says MHVA + PT "require a third of the time" of full finetuning. The supporting numbers are epochs to convergence: 14 against 49 on EK-100 (Table 4), which is 28.6%. Epochs are not wall-clock time, and the per-epoch cost differs between frozen and unfrozen backbones (backward passes still go through the frozen layers to reach the adapters). The GFLOPs column exists only in Table 1, with no figure for full finetuning, and it is not said whether it counts inference or training, or how many clips and crops per video.

3. Reproducibility

The datasets are public, and the backbones are standard released checkpoints, which is a good starting point. Beyond that, a reader could not rebuild the experiments from the text as rendered on Qeios. I could not find:

- A code or checkpoint link.
- Core hyperparameters of the method: the bottleneck dimension of MHVA, the number of window groups (heads), the prefix length, which Swin stages receive CAAs (§3.1 says cross-attention "performs well with different blocks", without saying which).
- Optimizer, learning rate, batch size, schedule, augmentation, input frames and resolution, and the hardware used.
- The feature extractors for audio, optical flow, and transcript on each dataset, and their output dimensions.

- Which splits were used. EK-100 has a validation set and a held-out test server; UDIVA v0.5 has only half the data released (§4.1), so the train/test partition matters; MPIIGI should name its protocol.
- Seeds and the number of runs per configuration.
- For the baselines the author re-ran (§4.2.1, Dual-Path and ST-Adapters "using their code"), the hyperparameter budget given to them compared with CM3T.

Equations 1 to 7 are partly illegible as rendered on Qeios: the math scrape breaks into single symbols, although the prose around them survives (for example, §3.2.2 gives 4.0 as the fixed value of the scaling factor that costs 1.1% on EK-100). I reviewed the method from the prose and Figure 2 caption, so I may have missed details that are in the typeset equations.

4. What specifically to fix

1. **Table 4 versus §4.2.3 and the abstract.** Reconcile 22.3% with the 28.9% shown for UDIVA and MPIIGI, and explain why one audio modality on MPIIGI costs 16.1% while two modalities on EK-100 cost 9.5%.
2. **§4.4.1, "only 5.8M additional parameters".** Show how 5.8M per modality fits the EK-100 total. A small table with parameters per module (MHVA, PT, each CAA, AdapterFusion) for each dataset would settle it.
3. Report mean and standard deviation over at least three seeds for the key rows of Tables 1 to 6, or at minimum for CM3T, MHVA + PT, and full finetuning. Without this, the sub-point differences (Table 3, 0.901 vs 0.899; Table 5, 48.2 vs 47.7; Table 6, 38.1 vs 37.4) cannot be read as wins.
4. **Tables 1 to 3, CLIP rows.** Add trainable-parameter counts for CM3T with CLIP-B and for FAT transformers, so the "5 times less" claim in §4.2.2 can be verified.
5. Add an implementation-details subsection (or appendix) with the items listed in section 3 above, and release code. I would put this first among the fixes for practical value.
6. **§3.2.2 and §4.3.** The ablations are reported only in prose: 1.1% for the learnable scale against a fixed value of 4.0, 0.7% for ReLU in prefix tuning, and "PT alone does not work very well" with no number. A single ablation table (MHVA alone, PT alone, MHVA + PT, fixed vs. learnable scale, tanh vs. ReLU, varying number of heads and bottleneck size) would make the design choices much more convincing.
7. **§4.4.1 and Table 1, GFLOPs column.** State whether GFLOPs are per clip at inference or per training step, give the full-finetuning figure, and replace "a third of the time" with measured wall-clock time

or GPU-hours.

8. **§4.2.1, re-run baselines.** Say how Dual-Path Adapters and ST-Adapters were tuned (search range, epochs) so the reader knows they were given a fair chance on the supervised ViT-B.
9. Small consistency points: Table 2 header says "Mean MSE" and Table 4 caption says "MSE"; Tables 2 and 3 write "CLIP" in some rows and "CLIP-B" in others; Table 5 has a stray % sign on 47.7%; Table 6 mentions Video Swin in the text but contains only ViViT-B rows. Also, note in the UDIVA table that lower is better.
10. **§4.2.2.** Fat transformers (ref. 38) share an author with this work. That is fine, but saying so in the text helps readers interpret the comparison, and the UDIVA result could be broken down per OCEAN trait.
11. **§5, limitations.** The limitation about dissimilar pretraining data is honest. One experiment would turn it into evidence: for example, the Something-Something v2 checkpoint mentioned in §1 applied to MPIIGL.

5. Recommendation

Minor revisions. The work makes a clear, useful point: adapters can be made to work on ordinary supervised video backbones, and the same mechanism extends to extra modalities with modest cost. The fixed-backbone comparison in Table 1 (39.8% against 28.7%) supports it well, and the experiments span three quite different datasets. The revisions I ask for are about making the efficiency numbers consistent (22.3% vs. 28.9%, the 5.8M per modality), reporting variance, and giving enough implementation detail for others to reuse the modules. None of this requires new methodology, and I expect the core conclusions to survive it.

Declarations

Potential competing interests: No potential competing interests to declare.