

Peer Review

Review of: "AVID: Adapting Video Diffusion Models to World Models"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper asks a practical question. If the best image-to-video diffusion models sit behind an API and their weights cannot be touched, can they still be turned into action-conditioned world models for robotics or games? The authors say yes, provided the API exposes the model's intermediate noise (or clean-sample) predictions and, for latent models, the encoder and decoder. Their method, AVID, trains a small 3D UNet adapter that receives the noisy video, the conditioning frame, and the pretrained model's prediction, conditions on actions through FiLM-style scale and shift, and outputs both its own prediction and a sigmoid mask. The final prediction is a mask-weighted mix of the two (Section 3.3, Eq. 5), trained with the ordinary denoising loss (Eq. 6).

Two further claims carry the paper. First, Section 3.2 argues that the product-of-experts composition of Yang et al. [23] gives biased samples because the noised marginal of a product of densities is not the product of the noised marginals, so the scores do not add. Second, on Coinrun (with a 97M pixel-space model the authors pretrained on the other 15 Progen games) and on RT1 (with the 1.4B DynamiCrafter), AVID matches or slightly beats ControlNet variants that need the weights and beats every baseline that, like AVID, sees only model outputs (Section 4.3, Figure 4, Tables 1 and 2).

2. Whether the conclusion holds on the data

I want to start with what I found genuinely good. The problem statement is new and useful, and the paper is honest about its boundary: Section 6 says plainly that AVID needs access to intermediate denoising outputs and the autoencoder, which most closed APIs do not offer today. The baseline set is unusually fair for an adapter paper. It includes parameter-matched from-scratch training, ControlNet-Small built to the same trainable budget, classifier guidance, action classifier-free guidance, the PoE

approach of [23], and the "cheating" upper references that need weights (full action finetuning, language finetuning, full ControlNet). Every method gets the same wall-clock budget. The Section 3.2 argument is short and correct, and the experiments then confirm it empirically. Table 1, row Product of Experts (71M), FVD 84.8, against AVID (71M) at 23.1. I also appreciated Section 6's admission that from-scratch training wins on Action Error Ratio in RT1, which is exactly the kind of sentence many papers leave out.

I checked the claims in Section 4.3 against the tables, one by one.

Coinrun 500k. "AVID performs the best for every evaluation metric at the smaller 22M model size" is correct: in Table 1, row AVID (22M), all six values (1.257, 28.5, 8.07, 0.666, 0.192, 22.4) beat the nearest competitor, ControlNet-Small (22M) (1.465, 34.7, 8.49, 0.652, 0.205, 21.9). At 71M, AVID has the best Action Error Ratio (1.154 vs 1.418 for ControlNet (71M)), while ControlNet wins the other five columns. The text says this accurately. One small point: the prose calls 22M the "smaller" size, but Table 1 labels that block "Medium" and has no Small block for Coinrun.

RT1. "AVID is the strongest in the metrics that make framewise comparisons (SSIM/LPIPS/PSNR) across all model sizes" holds in Table 2, with one tie: at Small, SSIM is 0.811 for both AVID (11M) and ControlNet-Small (10M). The paper also states that Action-Conditioned Diffusion is "consistently the best for Action Error Ratio," and that is right (2.238 vs 2.572 at Small, 1.737 vs 1.907 at Medium, 1.384 vs 1.609 at Large). It is worth adding that from-scratch training also has the lower FVD at Large (24.9 vs 39.3) and at Medium (36.7 vs 38.7). So in RT1, the picture is a trade-off: AVID wins on pixel fidelity to the ground truth, the from-scratch model wins on action consistency, and, at the two larger sizes, on FVD. The abstract's "outperforms existing baselines" rests on the normalized average in Figure 4. I would say that explicitly in the abstract because a reader who cares mainly about action following would read Table 2 differently.

The normalized score. I recomputed the bounds in Table 10 (RT1) and Table 11 (Coinrun) from Tables 1, 2, 4, and 5, and they are consistent with the method sets described in Appendix B.4: for example, the RT1 Action Error Ratio range 1.384 to 2.640 is Action-Conditioned Diffusion (145M) and ControlNet-Small (10M), and the Coinrun FVD minimum of 14.5 is AVID (71M) in Table 5. That is good bookkeeping. Still, min-max normalization over a hand-picked set of three methods makes the aggregate sensitive to which methods are in the set and to equal weighting of six correlated metrics (SSIM, LPIPS, and PSNR all measure framewise fidelity, so pixel accuracy effectively gets three votes out of six). A rank-based aggregate, or simply reporting the action-consistency metric and the fidelity metrics separately in Figure 4, would make the overall claim harder to dispute.

The ablations. This is the place where the text and the tables disagree most. Section 4.3 says that without the mask (NM) "performance across most metrics is worse for RT1, but similar for Coinrun500k". In Table 3, row No Mask (71M) is better than AVID (71M) on all six Coinrun metrics: 1.136 vs 1.154, 22.2 vs 23.1, 7.15 vs 7.33, 0.721 vs 0.713, 0.158 vs 0.161, 24.0 vs 23.8. In Table 6, at 22M, No Mask is better on five of six (FVD 28.6 vs 28.5 is the exception). The differences are small, so "similar" is defensible numerically, but the direction is consistent and goes against the mask. On RT1, the mask clearly helps (Table 6, Medium, Action Error Ratio 2.155 without the mask vs 1.907 with it). I would rewrite this paragraph to say that the learned mask helps on RT1 and gives no benefit on Coinrun, and offer a hypothesis; the paper's own Figure 4(d) discussion of how much each domain leans on the pretrained model would fit naturally here.

The No Conditioning ablation also needs a softer sentence. "Performance on most metrics gets significantly worse" is true in direction for most cells, but there are exceptions: Coinrun Large Action Error Ratio is 1.141 without conditioning vs 1.154 with it, RT1 Large FVD is 36.2 vs 39.3, and RT1 Medium PSNR is 25.1 vs 25.0 (Table 6). More importantly, "significantly" implies a test, and none is reported.

Variance. I could not find any mention of seeds, repeated runs, or confidence intervals. Each cell seems to be a single training run evaluated on 1024 test trajectories. With 1024 videos, FVD and FID estimates themselves have non-trivial sampling noise, and several conclusions rest on gaps of one to three percent (the bolding rule in Section 4.2 is "within 2% of the best"). Three seeds for AVID, ControlNet-Small, and Action-Conditioned Diffusion at one size per domain would tell the reader which of these gaps are real.

Hyperparameter selection. Appendix B.5 says that for tuned baselines the authors "choose the value that obtains the best FVD in our evaluation". If "our evaluation" is the 1024-trajectory test set, the baselines (PoE, the two guidance methods) were tuned on test data. That actually favors the baselines, not AVID, so it does not threaten the main result, but it should be stated clearly and ideally redone on a validation split.

Compute parity. All methods get the same wall clock (3 days on one A100 for Coinrun, 7 days on 4× A100 for RT1). AVID must also run a forward pass of the frozen 1.4B DynamiCrafter for every training example, which the from-scratch baseline does not. I could not tell from the text whether those forward passes come out of the same 7-day budget, or how many gradient steps each method completed. The same question applies at inference: AVID runs both models at each of the 50 DDIM steps (Table 9). Reporting steps, samples seen, and inference cost per video would make the efficiency comparison complete.

The "world model" claim. The title and the abstract frame AVID as a world model for decision-making. The evaluation is video prediction quality only; no planning or policy learning is run. Section 6 is candid

that this is future work. I would bring that caveat into the abstract.

3. Reproducibility

This part is strong. Code is linked (Section 4, the `microsoft/causica` repository, `research_experiments/avid`), though I did not run it and could not verify from the Qeios text what the repository contains. Both datasets are public; the Coinrun data generation is described precisely (Appendix B.1: 1000-step episodes from the first 100, 500, or 2500 levels under uniform random actions, evaluation on levels 10000 to 11000), and the RT1 split is given in counts (82,851 training episodes, 4,361 test episodes). Tables 8 and 9 list channels, heads, learning rates, EMA, batch size, noise schedule, sampler, and number of steps. The action-classifier and regression models for the Action Error Ratio are described with parameter counts (60M and 85M, Appendix B.3).

Gaps I noticed:

- No seeds are reported anywhere, for training or for sampling.
- DynamiCrafter is named and its resolution (320×512) is stated, but not the exact checkpoint. Several DynamiCrafter checkpoints exist, and naming the one used would close this.
- The 97M Procgan pretrained model is the authors' own (12 days on one A100, Appendix B.1). Whether its weights are released matters because the Coinrun results cannot be reproduced without it.
- The number of training steps per method within the wall-clock budget is not given.
- Classifier Guidance rows in Tables 1 and 2 have no parameter count, unlike all other rows.

4. What specifically to fix

1. **Section 4.3 and Table 3/Table 6.** Rewrite the No Mask sentence. On Coinrun 500k, No Mask is better than full AVID on all six metrics at 71M and on five of six at 22M; on RT1, the mask helps at every size. Say so, and discuss why the mask may only pay off when the base model is strong and backgrounds are static.
2. **Abstract and Section 6.** Qualify "outperforms existing baselines": in Table 2, Action-Conditioned Diffusion has the best Action Error Ratio at all three sizes and the lower FVD at Medium and Large. The claim is about the normalized average.
3. Add seeds and spread. Even three runs at one model size per domain (Tables 1 and 2) would do.
4. **Appendix B.5.** State which split was used for picking the PoE weight, the guidance weights, and the classifier-guidance w ; if it was the test set, move tuning to a validation split.

5. **Section 4, compute budget.** Report gradient steps or training samples completed per method, whether pretrained-model inference counts against AVID's budget, and sampling time per video.
6. **Section 4.2 vs Appendix B.3.** The main text defines Action Error Ratio as a ratio of errors (generated over real); the Coinrun formula in Appendix B.3 is a ratio of accuracies (real over generated). Both increase when generation gets worse, but they are different quantities, so one sentence aligning the definitions would help.
7. **Section 4.3, Table 1.** The text calls the 22M block "smaller" while the table labels it "Medium". Also, in Table 1, row Action Classifier-Free Guidance (22M), LPIPS is printed as 0.3495 (four decimals) while every other LPIPS entry has three.
8. **Reference [39].** FVD is attributed to "Cobbe K, Hesse C, Hilton J, Schulman J (2019)", the Progen authors. FVD is by Unterthiner et al.; please correct the author list.
9. **Figure 4 (a–d):** add per-metric panels or a supplementary table of normalized values, so the averages can be traced without redoing the arithmetic from Tables 10 and 11.
10. **Typos.** A handful, all easy: "a model the predicts" and "the question of to how to" (Section 1); "outputs of of" (contributions list); "without access it its weights" and "intial" (Section 3.2); "do not have to access to it's parameters" (Section 3.3); "trained on the trained on data" (Section 4.3); "conditoning" (Appendix B.2).
11. One downstream check, even small, would support the title: for instance, ranking a few candidate action sequences in RT1 with AVID rollouts and comparing against ground-truth outcomes. I understand this may be out of scope for this version.

5. Recommendation

Minor revisions. The problem is new, the analysis of why PoE composition is biased is correct, the baseline set is broad and fair, and the main numerical claims in Section 4.3 match Tables 1 and 2. What needs work is the reading of the mask ablation on Coinrun, the missing variance estimates, the test-set tuning question, and a more careful abstract about where AVID wins and where training from scratch still does better. None of these requires new methodology.

Declarations

Potential competing interests: No potential competing interests to declare.