MOST OF THE GAIN IS NOT VIDEO: TAKING APART A FINE-TUNED VIDEO QUESTION-ANSWERING SCORE

Anonymous authors
Paper under double-blind review

ABSTRACT

Video question-answering benchmarks rank video-language models, and fine-tuning on a benchmark's own training questions can lift a model's score by tens of points, a lift naturally read as better use of the video. A multiple-choice score can also rise for reasons unrelated to the video: a fine-tuned model learns the scorer's answer format, the answer statistics of the question set, and even answer letters the prompt never offers. Each effect is known in isolation, but their shares of a fine-tuned gain are rarely measured, and a leaderboard number cannot separate them. We measure them on UrbanVideo-Bench, where a LoRA adapter on Qwen2.5-VL-7B raises accuracy on a clip-disjoint development split from 29% to 71%. With prompt, scorer and split fixed and every comparison paired on the same predictions, a fifth to a quarter of the gain is the benchmark's released scoring rule misreading the zero-shot model's replies, and a seventh comes from questions whose answer letter the prompt never offers. An adapter trained on the question text alone, never seeing a frame, recovers 89 to 91% of the gain and scores within three points of each of three video-trained seeds; permuting the answer options costs the video-trained reference six points. Two further readings of the score shrink under controls: an apparent 20-point simulated-versus-real gap at least halves once both simulators are pooled, and three re-trained seeds place the 71% at the top of its recipe's range. These checks are cheap; we collect them in a reporting checklist for fine-tuned video-QA scores and provide the code and aggregate data.

1 INTRODUCTION

Many video-language benchmarks are reported zero-shot but ship thousands of public multiple-choice questions, and fine-tuning on them is the quickest way up a leaderboard; training on the test task is known to confound model comparisons (Dominguez-Olmedo et al., 2025). The resulting score is one number, read as better use of the video, yet it can also reflect learned answer statistics, compliance with the scorer's output format, and a favourable mix of question types and footage. Each ingredient is known in isolation. Models answer from question priors (Agrawal et al., 2018; Chen et al., 2024), prefer particular option identifiers (Zheng et al., 2024), and move with neutral changes of prompt format (Sclar et al., 2024). Much of an instruction-tuning gain can come from learning the output format (Kung & Peng, 2023), and question-only models trained on a video benchmark's own questions answer much of it (Ko et al., 2023; Xiao et al., 2024). How much of one fine-tuned gain each ingredient accounts for, measured on the same predictions, is rarely reported.

This work takes one such gain apart. UrbanVideo-Bench (Zhao et al., 2025a) asks multiple-choice questions about first-person drone video from two simulators and from real cities. Its paper reports zero-shot results for 17 models (best 45.5%) and one fine-tuning experiment, on sim-to-real transfer: LoRA (rank 128) on InternVL2-4B/8B trained on simulator questions raised accuracy on real footage from 28.3/26.5% to 31.5/31.7%. A plain LoRA adapter (Hu et al., 2021) on Qwen2.5-VL-7B (Bai et al., 2025), trained on answer letters from all three sources, reaches 70.9% on our clip-disjoint split (Figure 1), and the same procedure underlies the first-place entry in the UrbanCup 2026 video question-answering track (§4.5). To separate these ingredients, one evaluation setup governs every experiment, as in controlled design studies of embodied models (Wu et al., 2026)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

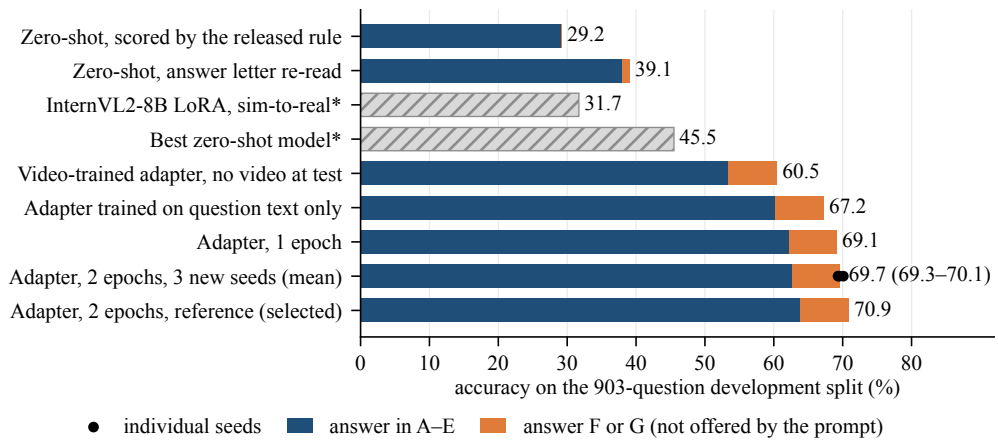


Figure 1: Accuracy on the 903-question development split, divided into points earned on questions whose answer is A–E (blue) and on the 65 questions whose answer is F or G (orange), letters the prompt never offers. The two zero-shot bars score the same development-file replies (16 frames of 640 pixels) with the benchmark's released rule and with the answer letter re-read. Adapter bars (passes of §2): development files (1 epoch, reference), prompt-ablation pass (no video at test), seed-and-probe pass (text-only adapter; three new seeds, dots). Hatched (∗): other models in the benchmark paper, for scale only; its LoRA bar is trained on simulator and tested on real footage.

(§2); training targets, input handling, and inference and capacity are varied mostly one at a time (§3), the score is decomposed by answer stratum, input, category and footage source (§4), and the findings become a reporting checklist (§5).

Three results change how the 70.9 reads. First, the benchmark's released scoring rule reads the zero-shot model's "Answer: C." as "A" (Figure 2), so 9.9 of the 41.6 points are output-format compliance (7.6 of 40.1 when the zero-shot model sees 12 frames). Another 6.0 come from 65 questions whose answer, F or G, lies outside the letter range the prompt offers. Second, an adapter trained on the question text alone scores 67.2% without ever seeing a frame, 2.1 to 2.9 points below each of three video-trained seeds (Figure 4 shows it on two items). Permuting the options costs every tuned model five to seven points with video and about half of its F/G answers (Figure 3). Third, an apparent 20-point simulated-versus-real gap compared one simulator with real footage; pooled over both simulators it is 9.7 points for the same checkpoint and 6.9 for the reference. The 70.9% is the top of its recipe's seed range ($69.66 \pm 0.40\%$).

Contributions. (i) An exact three-way decomposition of a fine-tuning gain into scorer compliance, answers the prompt cannot express, and the remainder, with a paired prompt ablation and an option-permutation test. (ii) Video/no-video controls and a text-only-trained adapter, three new seeds of the recipe, and paired tests for every training and input change with per-question predictions. (iii) Category and footage-source analyses: small categories swing by up to 19 points between near-identical checkpoints, and a 20-point simulated-versus-real gap belongs to one simulator, not to real footage. (iv) A reporting checklist. We propose no new method; the adapter has 80.7M trainable parameters and trains on one consumer GPU. This work does not aim to rank models on UrbanVideo-Bench, nor does it question the benchmark's published zero-shot results, which were extracted with a more lenient protocol than the released script (Zhao et al., 2025a, App. C.1). Rather, it asks what one fine-tuned score on it contains, and its checks apply to any benchmark that ships training questions.

Related work. MMBench counts a question as solved only if it is answered under every circular shift of the options (Liu et al., 2024). Choices also move with option order (Pezeshkpour & Hruschka, 2024) and with semantically neutral format changes (Rosenthal et al., 2026). Regular-expression answer extraction is a known source of evaluation error (Yu et al., 2025), and many image-benchmark questions are answerable without the image (Chen et al., 2024), as hypothesis-only baselines showed for natural language inference (Poliak et al., 2018). In video, an atemporal probe rivals full models (Buch et al., 2022), single frames or text often suffice (Lei et al., 2023;

Cores et al., 2025), and question-only models trained on a benchmark's own questions reproduce much of a fine-tuned model's accuracy and many of its predictions (Ko et al., 2023; Xiao et al., 2024). On UrbanVideo-Bench itself, a language model fine-tuned on the training questions without video was used to filter text-answerable test questions (Zhao et al., 2025b), and a mismatch between the prompt's letter range and the number of options was reported in the benchmark's public issue tracker (ZhanxyR, 2025). We use these known artefacts as instruments: what is new is how much of one in-distribution fine-tuning gain each accounts for, on the same predictions.

2 One evaluation setup for every experiment

Benchmark and split. The public release has 5,355 questions on 1,393 clips in sixteen categories; 319 questions (6.0%) have the answer F or G. With no official training split, we partition by clip (seed 42; whole clips assigned greedily until each category reaches 15%): 4,452 training questions on 1,149 clips and 903 development questions on 244 clips, with no clip on both sides. *The split is disjoint by clip, not by flight*: clip names encode a flight and a segment, and 215 of the 244 development clips come from flights that also contribute clips to training. Of 396 flights, 149 span both sides, and for 38 of them the whole-flight clip is in training while a segment is in development; both items in Figure 4 come from shared flights. On the 175 questions of the 29 flight-disjoint clips (Appendix C) the reference scores 73.7% (zero-shot 57.1%), the mean of three seeds 71.8% and the text-only adapter 68.0%. These three score $+3.5$ [-4.0, $+10.7$], $+2.7$ [-4.3, $+9.7$] and $+1.0$ [-6.9, $+8.9$] points higher there than on the other 728 questions, while the zero-shot model scores $+23.2$ [$+15.0$, $+31.3$] points higher: the subset differs in question mix (mostly recall and perception categories) rather than being harder for the adapters. The clip name also encodes the source: 413 development questions come from AerialVLN simulator clips, 305 from EmbodiedCity and 185 from real footage.

Prompt and scoring. The benchmark's released code, which we follow, sends the frames as separate JPEG images, then a fixed header and the question; the header, like the prompt printed in the benchmark paper, asks for "Option: []" with "one option from 'A' to 'E' ". Of the 903 questions 767 offer exactly A–E, 71 offer A–G and 65 offer A–C; the 65 questions with answer F or G (22 F, 43 G) necessarily lie among the 71 seven-option questions. The released scoring script reads the letter after "Option:" and otherwise takes the first character of the reply; our development parser applies the same rule. (The benchmark paper's own evaluation instead extracted answers with regular expressions and, for unstructured replies, GPT-4o over A–G. The released code sends up to 32 frames; we send 12, and 16 for the zero-shot development file.) Tuned models write "Option: [X]" in over 99% of replies, and both parsers give identical scores for every tuned pass. We therefore also re-read every reply with a lenient parser that additionally accepts "Answer: X" and a leading letter; it reads every reply in every pass and is used for every zero-shot comparison below.

Reference model. The reference is Qwen2.5-VL-7B-Instruct with LoRA on the attention and MLP projections of every language-model layer (rank 32, $\alpha = 64$; 80.7M parameters; vision encoder frozen), trained for two epochs on answer-only targets ("Option: [X]"). The training framework, ms-swift (Zhao et al., 2025c), feeds each clip as video input (at most 12 frames of at most 448×448 pixels, encoded as video tokens). Evaluation, as in the benchmark's code, sends 12 uniform JPEG frames of width 448 as separate images to vLLM (Kwon et al., 2023) at temperature zero, so training and evaluation see nearly the same frames in different encodings (Appendix G).

Evaluation passes and statistics. Numbers come from three passes with the same header, frames and parsers: the development prediction files (Tables 1, 2 and 4; Appendices A–B), a re-run of the prompt ablation (Table 3, Figure 5, Appendix Table 10) and a re-run for the seeds and probes (Figures 2–4; Appendix C, Tables 11–13). Figure 1 combines all three and names each bar's pass. Every comparison is made within one pass; across passes the same condition's accuracy differs by at most six questions (zero-shot with 12 frames 37.8 vs. 38.4%, reference without video 60.5 vs. 60.1%, with video 71.0 vs. 70.9%). Comparisons are paired on the same 903 questions (exact McNemar test; 10,000-sample paired bootstrap); evaluating the reference adapter twice changes whether 2 of 903 answers are correct, one each way. Values marked $*$ have no per-question predictions and carry no paired test. p-values are unadjusted. *Every training configuration except the reference recipe, which has three added seeds, is a single run.*

Table 1: Controlled study: every intervention against the reference (row A4, 640 of 903 correct; development files). Δq.: change in correct answers. *c/b*: questions the variant gets right and the reference wrong / the reverse; p: exact McNemar. $*$: no per-question predictions, so no paired test ($\approx$: known to one decimal). †: selected on this split. Training rows are single runs.

Row	Change from the reference	Dev %	Δq.	*c* / *b*	p
A4	reference: 7B LoRA, answer-only, 2 epochs, 12 frames at 448 px	70.87			
Training targets					
A2	1 epoch instead of 2 (evaluation frames not recorded)	69.10	−16	47 / 63	0.15
A6	self-generated rationales + 2× real footage	69.10	−16	37 / 53	0.11
A7	answer-only + 2× real footage, 2 epochs	69.10	−16	28 / 44	0.076
A14	preference optimisation on own errors (3,119 pairs)	69.66*	−11	–	–
Input handling					
A3	evaluate at 16 frames of 640 px instead of the training frame count	70.43	−4	25 / 29	0.68
A9	evaluate at 24 frames of 448 px	70.99	+1	18 / 17	1
A15	16 frames at 392 px, rank 16, seed 123	70.10	−7	38 / 45	0.51
A15	10 frames at 448 px, rank 64, seed 777	70.54*	−3	–	–
Inference and capacity					
–	same adapter, evaluation re-run (runtime noise)	70.87	+0	1 / 1	1
A8	3 samples at T=0.7, majority vote	67.0*	≈ −35	–	–
A10	32B, 4-bit QLoRA, same procedure	70.21*	−6	–	–
A11	7B/32B per-category router†	72.8*	≈ +17	–	–
A16	7B+32B log-probability fusion†	72.87*	+18	–	–
A16	best fusion of ≥ 3 of 5 members†	72.76*	+17	–	–
A11	oracle of 7B and 32B (upper bound)	78.07*	+65	–	–

3 A CONTROLLED STUDY ALONG THREE AXES

3.1 TRAINING TARGETS

Finding 1. *Answer-only fine-tuning carries the gain, and most of it arrives in the first epoch.* One epoch lifts the scored accuracy from 29.24% to 69.10% (Table 1, A2; the evaluation frames of that run were not recorded). The second epoch adds 12 questions evaluated at 16 frames (A3) and 16 at the training frame count (A4), which paired tests cannot separate from one epoch ($p = 0.15$ against A4). Against the re-read zero-shot model (39.09%) the reference gains 31.78 points (95% interval [28.13, 35.33]; 328 questions gained, 41 lost). On its own 4,452 training questions it scores 82.3%: it has not memorised them. Retraining the reference recipe with three further seeds on newer software (Appendix G) gives $68.11 \pm 1.55\%$ at the end of the first epoch (a mid-schedule checkpoint of the two-epoch run) and $69.66 \pm 0.40\%$ after two (mean ± s.d.; Figure 1, Appendix Table 11). The reference, selected on this split, lies 1.2 points above the two-epoch mean but within paired noise of every seed ($p \geq 0.15$); the second epoch helps one seed (+3.5 points, $p = 0.004$) and not the other two ($p = 0.21, 0.82$). Run-to-run variation is thus of the order of ± 1.5 points, and 70.9 is the top of what the recipe produces.

Finding 2. *Changing what the model is trained to output does not help.* Self-generated rationales with doubled real footage (A6) and doubled real footage alone (A7) both score 69.10% ($p = 0.11$, 0.076); preference optimisation (Hong et al., 2024) on 3,119 pairs mined from the reference's errors scores 69.66% ($*$). Doubling real footage moves real-footage accuracy from 65.4% to 67.0% (3 of 185 questions) and costs 8 and 11 questions (1.9 and 3.6 points) on AerialVLN and EmbodiedCity.

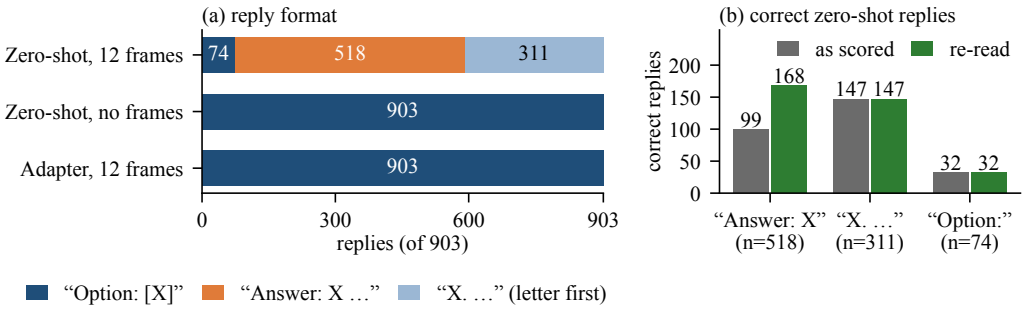


Figure 2: What the zero-shot model writes, and how it is scored (seed-and-probe pass). (a) Reply formats: with 12 frames, 518 of 903 zero-shot replies begin "Answer: X", which the released rule and our development parser read as "A"; without frames, and from the reference adapter, every reply contains "Option:". (b) Correct zero-shot replies with frames, by format, as scored and re-read: re-reading changes only the "Answer: X" replies; overall accuracy moves from 30.8% to 38.4%.

3.2 Input handling

Finding 3. *After fine-tuning, frame count barely matters.* Evaluating the reference with 24 instead of 12 frames changes the correctness of 35 answers, 18 gained and 17 lost ($p = 1$); an adapter trained with 16 frames at 392 pixels, rank 16 and another seed scores 70.10% ($p = 0.51$). Evaluating at 16 frames of 640 pixels instead of the training frame count changes four answers net ($p = 0.68$).

3.3 Inference and capacity

Finding 4. *Sampling and voting hurt an answer-only model.* Three samples at temperature 0.7 with a majority vote (Wang et al., 2023) score 67.0% ($*$), the largest loss in Table 1.

Finding 5. *A 4-bit 32B model (one run) does not move the score, but it errs differently.* Qwen2.5-VL-32B with 4-bit QLoRA (Dettmers et al., 2023) and the same procedure scores 70.21%. The two sizes agree on 78.7% of questions, and an oracle over the pair would score 78.07%. Fusing their answer-letter probabilities scores 72.87% (+2.0), the only gain in the study, but it was chosen on this split, and no fusion of three or more of five members exceeds 72.76% (all $*$).

4 Anatomy of one score

4.1 Scorer, answer letters and the prompt

Finding 6. *A fifth to a quarter of the headline gain is the scorer.* The zero-shot model rarely writes "Option:" when it sees frames: in the seed-and-probe pass, 74 of its 903 replies contain it, 518 begin "Answer: X" and 311 begin with the letter itself; without frames all 903 contain "Option:" (Figure 2). The released scoring rule, like our development parser, reads "Answer: C." as "A", so it credits every such reply whose gold letter is A, whatever the letter written (Figure 4, top). Reading the letter the model actually gave raises zero-shot accuracy in the development file from 29.24% to 39.09%, so 9.86 of the 41.64 points are output-format compliance that the adapter learned from its "Option: [X]" targets (Table 2). With 12 frames the share is 7.64 of 40.09 points (30.79% scored, 38.43% re-read). The penalty falls on the video condition only.

Finding 7. *Six points come from answers the prompt never offers, and about half of them follow the letter.* On the 65 F/G questions the re-read zero-shot model is right 10 times and the reference 64 times, 5.98 points of the gain. Sixty-five of the 71 seven-option questions have the answer F or G, so "answer G when seven options appear" alone would score 60.6% on them; Figure 4 (bottom) shows one such item. The remaining 25.80 points come from the 838 A–E questions (40.9% to 68.7%). Permuting each question's option contents (the answer's letter is re-drawn; it is unchanged for 166 questions) moves the reference from 70.87% to 64.78% (paired difference -6.1 $[-8.5, -3.7]$ points;

Table 2: The scored zero-shot-to-reference gain (development files), split exactly into the part due to the parser (zero-shot replies "Answer: X", scored as "A"), the 65 F/G questions and the 838 A–E questions. Zero-shot at 16 frames of 640 pixels (37.8–38.4% re-read at 12 frames of 448 pixels, Table 3 and Figure 3); the 9 F/G answers recovered by re-reading are counted as parser (the F/G share is 6.98 points if they are not).

Component	Questions	Accuracy change	Points of the gain
Parser: zero-shot writes "Answer: X"	903	29.2 → 39.1	9.86
Answer F/G (not offered by the prompt)	65	10/65 → 64/65	5.98
Answer A–E	838	40.9 → 68.7	25.80
Total (as scored)	903	29.2 → 70.9	41.64

Table 3: Paired prompt ablation (prompt-ablation pass): accuracy with the original A–E header and with a header listing each question's actual letter range, everything else fixed (12 frames at 448 pixels or no frames; replies re-read).

Model	Input	A–E header	Actual range	Δq.	p	F/G right (of 65)
Zero-shot	video	37.76	38.21	+4	0.45	9 → 10
Adapter	video	70.99	71.21	+2	0.69	63 → 63
Zero-shot	no video	28.57	29.13	+5	0.23	4 → 5
Adapter	no video	60.47	59.91	−5	0.062	64 → 62

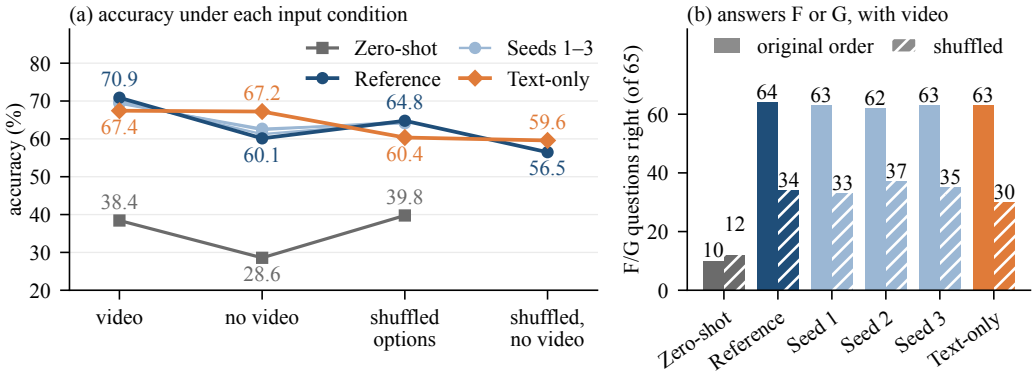
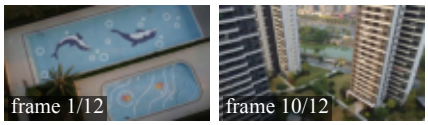


Figure 3: Input probes (seed-and-probe pass; 12 frames; zero-shot replies re-read). (a) Accuracy of each model with video, without video, with each question's option contents permuted (letters reassigned), and permuted without video (not run for the zero-shot model and the seeds). The text-only adapter is indifferent to frames; with video, every tuned model loses 5.0–7.1 points when the options are permuted, the zero-shot model none. (b) Correct answers, with video, on the 65 questions whose original answer is F or G, in the original and the permuted order.

94 answers lost, 39 gained). Of the 55 net losses, 32 are on the 71 seven-option questions, and with video every tuned model loses 5.0 to 7.1 points and about half of its F/G answers (Figure 3). On the 65 questions whose answer was F or G, the permuted reference still answers F or G 40 times although only 18 keep an F/G answer. It is right on 16 of those 18 and on 18 of the other 47, keeping 34 of its 64 correct answers there. Rebuilding the options without permuting them gives 70.54% ($p = 0.63$), so the rebuilt formatting is not the cause, and the zero-shot model is unaffected (38.43% to 39.76%, $p = 0.40$; Appendix Table 12).

Finding 8. *Neither model heeds the prompt's letter range.* In the pre-registered paired ablation (Table 3), replacing the A–E header by one listing each question's actual range changes the reference from 70.99% to 71.21% (+2 questions, $p = 0.69$; F/G 63/65 under both) and the zero-shot model from 37.76% to 38.21% (+4, $p = 0.45$; F/G 9 to 10). The instruction is not what kept the zero-shot model from answering F or G, and the adapter does not need to be told that F and G exist. Across the seven tuned prediction files F/G accuracy is 64/65 in six and 62/65 in the one-epoch file of the

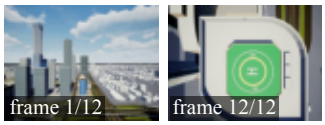


Zero-shot, 12 frames: "Answer: E. I am currently over a road, …"
as scored A ✓ re-read E ✗ no frames D ✗
Reference adapter: 12 frames A ✓ no frames E ✗
Text-only adapter: 12 frames A ✓ no frames A ✓
Seeds 1–3, 12 frames: A ✓ A ✓ A ✓

Landmark Position · real footage · gold answer A

Q: At the initial position, you are asked to navigate towards the outer edge of the community and make a half circle. When you reach the current location, what is your direction to the destination?

A. I am currently over the residential area, close to the outer edge, and ready to hover.
B. I am currently far away from residential areas and over a road.
C. I am currently in the center of the residential area and above the park.
D. I am currently outside the residential area and have completed half a circle.
E. I am currently over a road, facing away from a residential area.



Zero-shot, 12 frames: "Answer: C. I fly forward. Reason: The …"
as scored A ✗ re-read C ✗ no frames C ✗
Reference adapter: 12 frames G ✓ no frames G ✓
Text-only adapter: 12 frames G ✓ no frames G ✓
Seeds 1–3, 12 frames: G ✓ G ✓ G ✓

Action Generation · EmbodiedCity simulator footage · gold answer G

Q: At the initial position, you are asked to navigate to the center of the tarmac. What is your next move, given the target location for your next move?

A. I rise.
B. I rotate the camera downward.
C. I fly forward.
D. I turn left.
E. I turn right.
F. I rotate the camera upward.
G. I descend.

Figure 4: Two development items (UrbanVideo-Bench, MIT licence) under one view: two of the 12 frames each model receives, the question and options (gold in green; F and G, which the "A to E" header never offers, in brown), and every system's answer in the seed-and-probe pass (✓ right, ✗ wrong). Top: chosen by a fixed rule as an item that needs the video (the reference is right with frames and wrong without them, and the zero-shot model misses it; Appendix E); the adapter trained on question text alone answers it anyway, and the released rule scores the zero-shot reply "Answer: E." as "A", the gold letter. Bottom: the first development question, by id, whose answer the header cannot express. Both clips come from flights that also have clips in training (§2).

reference recipe, so the differences between tuned checkpoints (69.10–70.99%) lie almost entirely in the A–E questions (Appendix Table 6).

4.2 How much of the score needs the video

Finding 9. *Most of what fine-tuning buys is available without the video.* Given only the question text, the video-trained adapter scores 60.47% and answers 64 of the 65 F/G questions (prompt-ablation pass; Appendix Table 10). Fine-tuning adds 31.9 points without video and 33.2 with it; video adds 9.2 points to the zero-shot model and 10.5 to the adapter. Without video the adapter scores 52.5% on AerialVLN, 73.4% on EmbodiedCity and 56.8% on real footage, and it already exceeds the zero-shot model with video in 12 of 16 categories. Video adds the most to it on Start/End Position (+39.1), Scene Recall (+30.4) and Goal Detection (+24.1), and almost nothing on Action Generation, the largest category (+1.5, 197 questions; Figure 5). Because this adapter never saw text-only inputs in training, 60.5% understates what the text supports. An adapter trained with the same recipe on the question text alone scores 67.22% without video (seed-and-probe pass; Figure 3), 7.1 points above the video-trained reference evaluated without video in that pass ($p = 4 \times 10^{-7}$). It is 2.1, 2.9 and 2.3 points below the three video-trained seeds with video ($p = 0.13, 0.04, 0.10$; mean 2.4) and 3.7 below the development-selected reference ($p = 0.009$). Adding frames at test time does not help it (67.44%, $p = 0.92$), and it answers 63/65 F/G questions; it even answers the item of Figure 4 (top), which was chosen because the reference needs the video for it. In that pass it thus recovers 91% of the reference's gain over the zero-shot model with frames as scored by the released rule, and 89% with replies re-read; training with video adds 2.4–3.7 points. The video-trained seeds do use the frames (video adds 6.6–8.9 points per seed, all $p < 10^{-5}$), but they learned to answer largely from the question text: its templates, option lists and answer statistics. This holds on the flight-disjoint subset, where the text-only adapter is 3.8 points below the seed mean (95% interval -2.1 to 9.7; Appendix C). Because training encodes clips as video tokens while evaluation sends separate images, the video-trained adapters may also use frames less than they could.

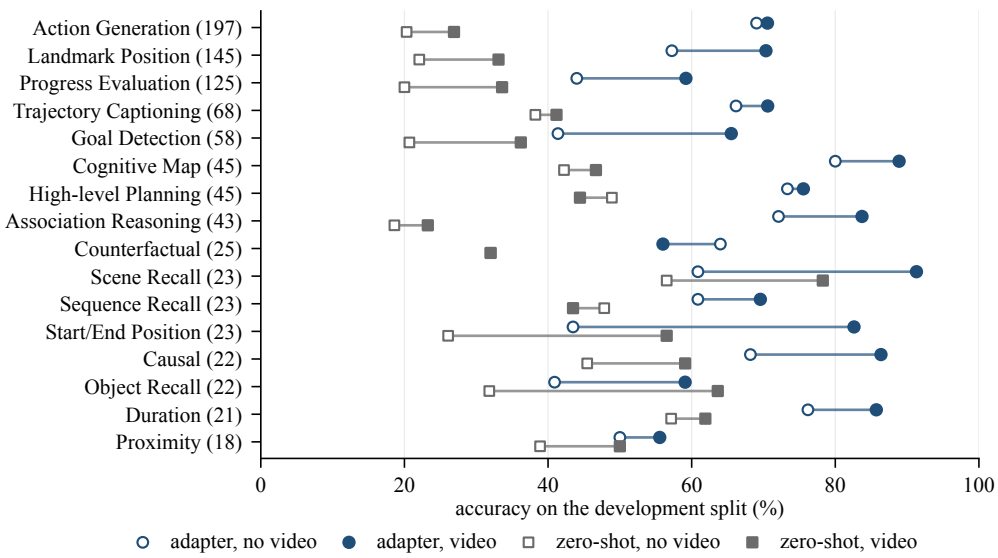


Figure 5: Per-category accuracy without video (open markers) and with the 12 frames (filled), for the reference adapter (circles) and the zero-shot model (squares); prompt-ablation pass, A–E header, re-read replies. Category size in parentheses.

Table 4: Accuracy by footage source (development files). Comparing EmbodiedCity alone with real footage in the A3 row (82.6 vs. 62.7%) gives a 20-point gap; pooling both simulators, 9.7.

Evaluation	AerialVLN (n=413)	EmbodiedCity (n=305)	Both sim. (n=718)	Real (n=185)
Zero-shot, re-read	39.5	37.7	38.7	40.5
1 epoch (A2)	63.2	80.0	70.3	64.3
Rationales + 2× real (A6)	61.7	80.3	69.6	67.0
2× real footage (A7)	62.2	79.7	69.6	67.0
At 16 frames, 640 px (A3)	64.9	82.6	72.4	62.7
Reference (A4)	64.2	83.3	72.3	65.4
Gain over zero-shot	+24.7	+45.6	+33.6	+24.9

4.3 Question Category

Finding 10. *The gain concentrates in three large categories, and small categories cannot be ranked.* Action Generation (197), Landmark Position (145) and Progress Evaluation (125) make up 51.7% of the split and supply 18.94 of the 31.78 points gained over the re-read zero-shot model; the mean over sixteen categories rises by 25.8 points (47.4% to 73.2%). Across the seven tuned files, whose overall accuracies lie within 1.9 points, categories with fewer than 30 questions move by up to 19.0 points (Duration, 71.4–90.5%). Object Recall, the only category where the reference is below zero-shot (two questions of 22), is below it in four tuned files and at or above it in three. We do not claim that fine-tuning hurts any category (Appendix Figure 6 and Table 8).

4.4 Footage Source

Finding 11. *Pooled over both simulators the simulated-versus-real gap is 3 to 10 points, not 20, and the gain on real footage matches the larger simulator.* Comparing EmbodiedCity alone with real footage gives 82.6% against 62.7% for the checkpoint evaluated at 16 frames (Table 4), a 20-point gap, and real-footage Progress Evaluation scores 0 of 8 questions there (1 of 8 for the reference). Pooling both simulators, that checkpoint scores 72.4% on simulated footage (gap 9.7) and the reference 72.3% against 65.4% (6.9); across the tuned files of Table 4 the pooled gap is 2.6 to 9.7 points. AerialVLN, the larger simulator, scores 64.2% with the reference, below real footage. The re-read zero-shot model is at 38.7% on simulated and 40.5% on real footage. Fine-tuning adds

24.7 points on AerialVLN, 24.9 on real footage and 45.6 on EmbodiedCity, which holds 51 of the 65 F/G questions (80.3% on its A–E questions). The outlier is one simulator, not real footage.

4.5 CHALLENGE RESULT

The procedure underlies the first-place entry in the UrbanCup 2026 video question-answering track; the organisers' hidden-set score for that entry is 1.5 points below the reference's development accuracy. The entry combined the 7B reference with the 32B model (72.87% on our split); the organisers' notice does not name the scored submission and the hidden set's composition is not documented, so we do not decompose that score.

5 A REPORTING CHECKLIST FOR FINE-TUNED VIDEO-QA SCORES

What this case study suggests reporting alongside any fine-tuned multiple-choice video-QA score:

1. **Split by recording, not by clip**: keep all clips of a flight on one side, and publish the clip ids of each side with their hash.
2. **Re-read zero-shot replies with a lenient parser** and report both numbers; a baseline that fails the output format is not a baseline (Finding 6).
3. **List each question's actual options**, report option-count and answer-letter distributions, and report accuracy on questions the prompt cannot express (Findings 7–8).
4. **Report a no-video control and a text-only-trained adapter**, and evaluate with permuted option order (Findings 7 and 9).
5. **Report gains by category with** n, micro and macro, and the spread across seeds or near-equivalent checkpoints (Finding 10).
6. **Report accuracy per footage source**, not per "sim" vs. "real" bucket (Finding 11).
7. **Use paired tests** with discordant counts, re-evaluate once to measure runtime noise, report seeds, and disclose selection on the evaluation split (Findings 1–5).

6 LIMITATIONS AND CONCLUSION

This is a case study: one benchmark, one model family and one split, which served for all development decisions; no untouched test set is reported. The split separates clips but not flights. On the 175-question flight-disjoint subset the reference, the seed mean and the text-only adapter each score at least as high as on the rest, but that subset is small and differs in question mix. Training encodes clips as video tokens while evaluation, following the benchmark's code, sends separate images, so the video-trained adapters may use frames less than they could. Only the reference recipe has seeds; the other training configurations, including the text-only adapter, are single runs, and paired tests measure evaluation differences between fixed checkpoints. In the main text two effects have $0.001 < p < 0.01$ (the seed-2 epoch effect and the text-only adapter's gap to the selected reference) and every other effect called significant has $p \leq 3 \times 10^{-6}$; tests on the flight-disjoint subset are weaker (Appendix C). The lenient parser is ours; it changes no tuned-model score and reads every reply in every pass. Rows A8, A10, A11, A14, A15 (second adapter) and A16, and the A5 simulated/real figure, have no per-question predictions.

In this work, we took one fine-tuned video-QA score apart and found that it is not one quantity. On UrbanVideo-Bench a fifth to a quarter of the gain is the scorer, and a seventh comes from answers the prompt never offers. An adapter that never sees a frame gets within 2.4 points of the video-trained recipe's mean, the apparent 20-point sim-to-real gap is one easy simulator, and the 70.9% is the top of its recipe's seed range; neither scale nor training or decoding changes add to a single adapter. Re-reading replies, text-only controls, permuted options and reporting by stratum, category and source are cheap, and each makes one of these visible; we hope the checklist helps readers of fine-tuned leaderboard scores ask what each number contains.

ETHICS STATEMENT

All data are the public UrbanVideo-Bench release (MIT licence); we collected no data from people, and the frames shown in the figures come from that release. The work evaluates existing models and proposes no deployable system.

REPRODUCIBILITY STATEMENT

The supplementary material contains the analysis, re-scoring, statistics and plotting code that regenerates every table and figure from the aggregate counts; those aggregate counts, with the hashes of the ground truth, the lists of training and development clips and every development prediction file; the question and clip ids of both sides of the split; the frozen prompt-ablation scripts; and the training and evaluation scripts of the seed-and-probe runs. Per-question correctness vectors (question id and 0/1), which the paired tests and bootstrap intervals need, will be released with the camera-ready version. Training hyperparameters and software versions are in Appendix G; questions and videos are those of the public UrbanVideo-Bench release. Every training configuration except the reference recipe (three added seeds) is a single run, and Table 1 marks the values that have no per-question predictions.

AI USE STATEMENT

This work was carried out with LLM-based coding agents (Anthropic Claude models) under the author's direction. In the categories of the ICLR 2027 policy, the agents implemented methods (the analysis, re-scoring, statistics and plotting code, and the training and evaluation scripts of the seed, permuted-option and text-only runs); reformatted the public benchmark data into training and evaluation inputs; designed and gave feedback on experiments (they proposed the lenient re-read, the no-video control, the text-only adapter, the option-permutation probe and the seed runs, and the author approved each batch before it ran); ran the experiments (they launched and monitored the GPU runs on a lab machine and on rented machines); interpreted results and suggested experimental parameters; created and modified the figures; searched for, summarised and checked the related literature, including a search for prior work that would make the study redundant, and formatted the references; suggested the structure of the paper and its keywords; and drafted and edited its text. Generative AI was not used to generate synthetic data or to create, alter or impute any result: every number is computed by scripts from recorded model predictions, and the questions, answers and frames are those of the public benchmark. AI-assisted work was checked by agent audits that recomputed every number from the per-question records and every reference from its primary source, and by the author's review of the design, results and text. The author takes full responsibility for the paper.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. arXiv:1712.00377.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2206.01720.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, et al. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. arXiv:2403.20330.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in time: A new temporal benchmark for VideoLLMs. In *British Machine Vision Conference (BMVC)*, 2025. arXiv:2410.07752.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2305.14314.

Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.07890.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11170–11189, 2024. arXiv:2403.07691.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J. Kim. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Po-Nien Kung and Nanyun Peng. Do models really learn to follow instructions? an empirical study of instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 2: Short Papers*, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, et al. Efficient memory management for large language model serving with PagedAttention. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2023. arXiv:2309.06180.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 487–507, 2023. arXiv:2206.03428.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, et al. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2307.06281.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024. arXiv:2308.11483.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018. arXiv:1805.01042.

Fabio Rosenthal, Sebastian Schmidt, Thorsten Graf, Thorsten Bagdonat, Stephan Günnemann, and Leo Schwinn. Unexplored flaws in multiple-choice VQA make benchmarking unreliable. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026. arXiv:2511.22341.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11324.

Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2203.11171.

Xiao-Ming Wu, Bin Fan, Kang Liao, Jian-Jian Jiang, Runze Yang, Yihang Luo, Zhonghua Wu, Wei-Shi Zheng, and Chen Change Loy. VLANeXt: Recipes for building strong VLA models. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.18532.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? visually grounded video question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2309.01327.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. xFinder: Large language models as automated evaluators for reliable evaluation. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2405.11874.

ZhanxyR. VLMEvalkit support and several issues. GitHub issue #1 in the UrbanVideo-Bench code repository, opened 23 May 2025, 2025. URL `https://github.com/EmbodiedCity/UrbanVideo-Bench.code/issues/1`.

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and Yong Li. UrbanVideo-Bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 32400–32423, 2025a. arXiv:2503.06157.

Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. Embodied-R: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. *arXiv preprint arXiv:2504.12680*, 2025b.

Yuze Zhao, Jintao Huang, Jinghan Hu, et al. SWIFT: A scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence (Demonstration Track)*, volume 39, pp. 29733–29735, 2025c. arXiv:2408.05517.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.03882.

A Provenance of every number

Every number in the paper is computed from aggregate counts of per-question predictions: for the development evaluations and the prompt ablation on the machine that holds the questions, videos and predictions, from which only counts left; for the seed-and-probe runs from their per-question outputs. Table 5 describes every evaluation in the paper; Table 6 breaks each development evaluation down by answer stratum; Table 7 gives the paired test of each against the reference. The development evaluations of rows A6 and A7 are matched to those rows by their training targets (self-generated rationales with doubled real footage for A6) and because they reproduce the recorded 69.1% overall and 65.4→67.0% on real footage; one further evaluation is the first epoch of A7. The first A15 adapter scores 70.10% where the development record has 69.99% (one question). Values without a prediction file (A8, A10, A11, A14, the second A15 adapter and A16, and the A5 simulated/real figure) are transcribed from the development record with their row ids. File names and checksums of all evaluations are listed in the supplementary material.

B Per-category accuracy

Table 5: Every evaluation in the paper: model and training, frames at test and correct answers of 903 questions. All use the A–E header of the benchmark's code; the prompt ablation's actual-range header is in Table 3.

Evaluation (row of Table 1)	Frames at test	Correct of 903
Development evaluations (A–E header; replies scored by the released rule)		
Zero-shot Qwen2.5-VL-7B; re-read in brackets	16 at 640 px	264 (353)
Adapter, answer-only, 1 epoch (A2)	not recorded	624
Adapter, self-generated rationales + 2× real footage (A6)	not recorded	624
Adapter, answer-only + 2× real footage, epoch 1 of A7	not recorded	625
Adapter, answer-only + 2× real footage, 2 epochs (A7)	not recorded	624
Reference adapter, 2 epochs (A3)	16 at 640 px	636
Reference adapter, 2 epochs (A4)	12 at 448 px	640
Reference adapter, 2 epochs (A9)	24 at 448 px	641
Reference adapter, evaluation re-run with answer-letter probabilities	not recorded	640
Adapter trained at 16 frames of 392 px, rank 16, seed 123 (A15)	not recorded	633
Prompt-ablation pass (A–E header; replies re-read; with frames / without)		
Zero-shot Qwen2.5-VL-7B	12 at 448 px / none	341 / 258
Reference adapter	12 at 448 px / none	641 / 546
Seed-and-probe pass (A–E header; replies re-read; with frames / without)		
Zero-shot Qwen2.5-VL-7B	12 at 448 px / none	347 / 258
Reference adapter	12 at 448 px / none	640 / 543
Reference recipe retrained, seeds 1, 2, 3 (2 epochs)	12 at 448 px	626, 633, 628
Adapter trained on question text only (2 epochs)	12 at 448 px / none	609 / 607

Table 6: Every development evaluation by answer stratum; last column: change in correct answers from the reference.

Evaluation (row)	All (903)	A–E (838)	F–G (65)	Δ vs. ref.
Zero-shot, 16 frames of 640 px, as scored	29.24	31.38	1/65	-376
Zero-shot, same outputs re-read	39.09	40.93	10/65	-287
Adapter, 1 epoch (A2; evaluation frames not recorded)	69.10	67.06	62/65	-16
Rationales + 2x real footage (A6)	69.10	66.83	64/65	-16
2x real footage, epoch 1 (A7)	69.21	66.95	64/65	-15
2x real footage, epoch 2 (A7)	69.10	66.83	64/65	-16
Adapter, 2 epochs, 16 frames of 640 px (A3)	70.43	68.26	64/65	-4
Adapter, 2 epochs, 12 frames of 448 px (A4, reference)	70.87	68.74	64/65	$+0$
Same adapter, 24 frames of 448 px (A9)	70.99	68.85	64/65	$+1$

Table 7: Paired comparison of every development evaluation with the reference (c: this evaluation right, reference wrong; b: the reverse, as in Table 1; interval: 10,000-sample paired bootstrap of this evaluation minus the reference).

Evaluation (row)	Dev %	c	b	McNemar p	95% CI of Δ (pp)
Zero-shot, re-read	39.09	41	328	1×10^{-56}	$[-35.33, -28.13]$
1 epoch (A2)	69.10	47	63	0.15	$[-3.99, +0.55]$
Rationales + 2× real (A6)	69.10	37	53	0.11	$[-3.88, +0.22]$
2× real footage, epoch 1 (A7)	69.21	33	48	0.12	$[-3.65, +0.33]$
2× real footage (A7)	69.10	28	44	0.076	$[-3.65, +0.00]$
At 16 frames, 640 px (A3)	70.43	25	29	0.68	$[-1.99, +1.22]$
Reference (A4)	70.87	–	–	–	reference
Reference, re-run with log-probs	70.87	1	1	1	$[-0.33, +0.33]$
Reference, 24 frames of 448 px (A9)	70.99	18	17	1	$[-1.22, +1.33]$
16 frames at 392 px, rank 16, seed 123 (A15)	70.10	38	45	0.51	$[-2.77, +1.22]$

Table 8: Per-category accuracy. "Min–max" is over the seven tuned development evaluations; "Tuned < zero-shot" counts evaluations strictly below zero-shot; "Share of gain" is the category's contribution to the 31.78-point gain over the re-read zero-shot model (the scored zero-shot model is 40.1% macro).

Category	n	Zero-shot	Ref.	Min–max over 7 tuned	Tuned < zero-shot	Share of gain (pp)
Association Reasoning	43	27.9	81.4	76.7–83.7	0/7	2.55
Action Generation	197	25.9	71.1	69.5–71.1	0/7	9.86
Cognitive Map	45	48.9	88.9	86.7–91.1	0/7	1.99
Landmark Position	145	31.0	69.7	62.8–71.0	0/7	6.20
Duration	21	57.1	90.5	71.4–90.5	0/7	0.78
Start/End Position	23	56.5	87.0	69.6–87.0	0/7	0.78
Trajectory Captioning	68	42.6	72.1	72.1–77.9	0/7	2.21
Causal	22	59.1	86.4	72.7–90.9	0/7	0.66
High-level Planning	45	51.1	75.6	68.9–77.8	0/7	1.22
Goal Detection	58	44.8	67.2	55.2–69.0	0/7	1.44
Sequence Recall	23	47.8	69.6	56.5–69.6	0/7	0.55
Progress Evaluation	125	36.8	57.6	57.6–60.0	0/7	2.88
Counterfactual	25	36.0	52.0	40.0–52.0	0/7	0.44
Scene Recall	23	73.9	87.0	73.9–87.0	0/7	0.33
Proximity	18	50.0	55.6	55.6–61.1	0/7	0.11
Object Recall	22	68.2	59.1	59.1–77.3	4/7	−0.22
All (micro)	903	39.1	70.9	69.1–71.0		31.78
Mean of categories (macro)	16	47.4	73.2			

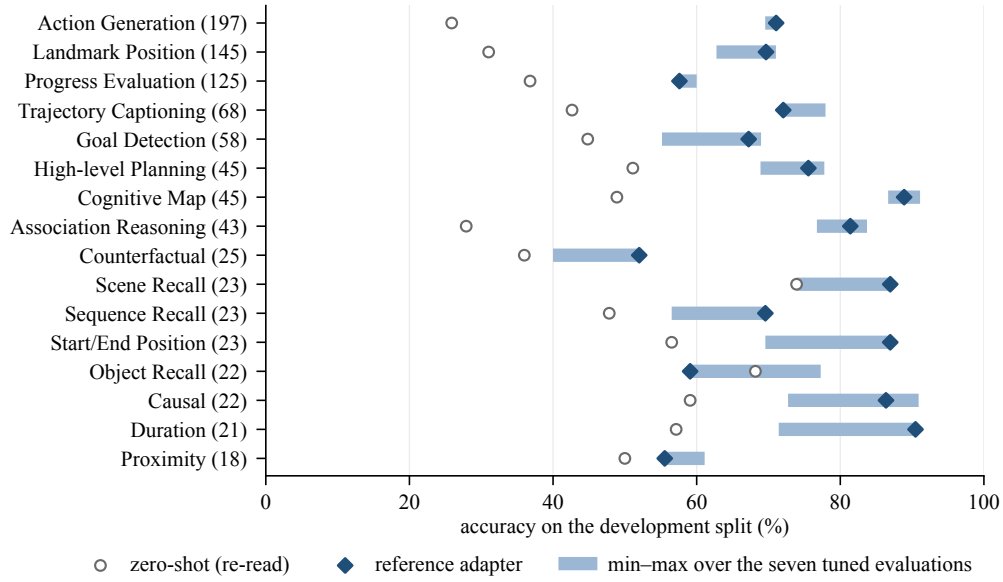


Figure 6: Per-category re-read zero-shot accuracy (circle), reference (diamond) and range over the seven tuned development evaluations (bar); category size in parentheses.

C FLIGHT-DISJOINT SUBSET

The 29 development clips whose flight has no clip in training carry 175 questions (100 AerialVLN, 64 real, 11 EmbodiedCity; 4 with answer F or G). Table 9 re-scores the seed-and-probe pass on them (zero-shot replies re-read). Between arms on this subset, the mean of the three seeds with video is $+3.8$ [-2.1, $+9.7$] points above the text-only adapter without video, and the reference with video $+5.7$ [-1.1, $+12.6$] above it ($p = 0.13$); video adds $+13.7$ [$+6.9$, $+20.6$] points to the reference, $+8.8$ [$+3.0$, $+14.7$] to the seed mean and $+1.7$ [-3.4, $+6.9$] to the text-only adapter. Subset minus the other 728 questions: zero-shot $+23.2$ [$+15.0$, $+31.3$], reference $+3.5$ [-4.0, $+10.7$], seed mean $+2.7$ [-4.3, $+9.7$], text-only adapter without video $+1.0$ [-6.9, $+8.9$] points (independent-groups bootstrap). In the development files the reference scores 74.3% on the subset and the re-read zero-shot model 55.4% (gain $+18.9$ [$+11.4$, $+26.3$] points).

Table 9: Accuracy on the flight-disjoint subset (seed-and-probe pass). Δ: paired difference from the zero-shot model with the same input (video rows: zero-shot with video; no-video rows: zero-shot without video), 10,000-sample paired bootstrap; p: exact McNemar (not defined for the seed mean). Single evaluation per row; seeds are single training runs.

Model and input	Subset (n=175)	Rest (n=728)	All (903)	Δ (subset)	p
Zero-shot, video (re-read)	57.1	33.9	38.4	–	–
Zero-shot, no video (re-read)	45.1	24.6	28.6	–	–
Reference, video	73.7	70.2	70.9	$+16.6$ [$+9.1$, $+23.4$]	2.5×10^{-5}
Reference, no video	60.0	60.2	60.1	$+14.9$ [$+7.4$, $+22.3$]	0.00031
Seed 1, video	72.6	68.5	69.3	$+15.4$ [$+8.0$, $+22.9$]	0.00014
Seed 2, video	71.4	69.8	70.1	$+14.3$ [$+6.3$, $+21.7$]	0.00062
Seed 3, video	71.4	69.1	69.5	$+14.3$ [$+6.9$, $+21.7$]	0.00047
Mean of seeds, video	71.8	69.1	69.7	$+14.7$ [$+7.4$, $+21.7$]	–
Mean of seeds, no video	63.0	61.9	62.1	–	–
Text-only adapter, no video	68.0	67.0	67.2	$+22.9$ [$+14.9$, $+30.9$]	1.6×10^{-7}
Text-only adapter, video	69.7	66.9	67.4	$+12.6$ [$+4.6$, $+20.0$]	0.0026

D SEED, PERMUTED-OPTION, TEXT-ONLY AND VIDEO/NO-VIDEO TABLES

Table 10: Accuracy with and without the 12 video frames (prompt-ablation pass: same frames and questions; A–E header; re-read replies), with paired 95% bootstrap intervals of each difference.

	No video	Video (12 frames)	Video adds
Zero-shot	28.57	37.76	$+9.2$ [6.0, 12.5]
Adapter	60.47	70.99	$+10.5$ [7.8, 13.4]
Fine-tuning adds	$+31.9$ [28.2, 35.7]	$+33.2$ [29.6, 36.8]	

Table 11: Seed replication of the reference recipe on the development split (seed-and-probe pass; each seed one training run). Seed epoch-1 values are mid-schedule checkpoints of the two-epoch run. Reference row: epoch-1 value from the development run (evaluation frames not recorded), other columns from the same pass as the seeds. Shuffled: option contents permuted per question, letters reassigned.

Run	1 epoch	2 epochs	2 ep., no video	2 ep., shuffled options	F/G (2 ep.)
Seed 1	69.66	69.32	62.68	64.12	63/65
Seed 2	66.56	70.10	61.24	64.34	62/65
Seed 3	68.11	69.55	62.35	64.56	63/65
Mean $\pm$ s.d.	68.11 ± 1.55	69.66 ± 0.40	62.09 ± 0.75	64.34 ± 0.22	
Reference (epoch 1: development run)	69.10	70.87	60.13	64.78	64/65

Table 12: Permuted ("shuffled") options (seed-and-probe pass): option contents permuted with a fixed per-question seed; "identity rebuild" re-renders the options unpermuted. F/G and 7-option columns count correct answers on the questions whose original answer is F/G and that offer seven options.

Model	Options	All	F/G-answer q. (65)	7-option q. (71)	No video
Zero-shot	original	38.43	10	12	28.57
	shuffled	39.76	12	12	–
Adapter (reference)	original	70.87	64	67	60.13
	shuffled	64.78	34	35	56.48
	identity rebuild	70.54	64	67	

Table 13: An adapter trained on question text only against the reference adapter trained with video (seed-and-probe pass).

Adapter trained on	Eval. no video	Eval. video	Shuffled, no video	Shuffled, video
Question text only	67.22	67.44	59.58	60.35
Video (reference)	60.13	70.87	56.48	64.78

E EXAMPLE ITEMS

The two items in Figure 4 are public UrbanVideo-Bench development questions (MIT licence); their frames were extracted with the evaluation sampler (12 uniform frames; two shown, resized to width 640), and every answer shown comes from the seed-and-probe pass. The top item was chosen by a fixed rule on the development files and the prompt-ablation pass: among five-option development questions whose text, options included, is shorter than 700 characters, that the reference adapter answers correctly with video but not without it, and that the zero-shot model with video misses (26 such questions; 13 AerialVLN, 7 EmbodiedCity, 6 real), take the first by question id in the first non-empty (source, category) pair, sources ordered real $>$ AerialVLN $>$ EmbodiedCity and categories Landmark Position $>$ Progress Evaluation $>$ Cognitive Map $>$ Action Generation $>$ Start/End Position. It relaxes a stricter rule (also requiring the zero-shot model without video to be wrong, with the length capped at 420 characters), which leaves only two candidates. The bottom item is the first development question, by question id, whose answer is F or G.

F PROMPT ABLATION DETAILS

The prompt ablation is pre-registered: its code was frozen before it ran, and its tuned-model arm ran unchanged. Both arms use the reference adapter (its weights checked against the development copy), the same 12 JPEG frames per video (the encoded payload is built once and reused), temperature zero and counterbalanced arm order by question id. The zero-shot and text-only arms were added for this paper with a copy of the code that changes only the served model (base model without adapter) or removes the images; the frozen code was not modified. All four runs used one otherwise idle 48 GB GPU for about 32 minutes in total. The frozen scorer reads only "Option: [X]"; it leaves 821 of 903 zero-shot replies unparsed, so zero-shot arms are scored with the lenient parser of §2 (tuned-arm scores are identical under both). In the tuned arm the seven-option questions score 66/71 under both headers, the three-option questions 43 and 46 of 65, and 2 of the 767 questions whose header is identical change their answer (runtime noise); the frozen scorer's paired 95% interval for the tuned arm is [-0.33, +0.78] points. The tuned arm with the A–E header reproduces the development evaluation of the reference to within one question (70.99% vs. 70.87%). Without video the adapter answers 68 of the 71 seven-option questions, with video 66.

G TRAINING AND EVALUATION DETAILS

LoRA rank 32, $\alpha = 64$, dropout 0.05, on the query, key, value, output, gate, up and down projections of every language-model layer (80,740,352 parameters); vision encoder frozen; bf16; learning rate 10^{-4} with 3% warm-up; per-device batch 1 with gradient accumulation 8; maximum length 10,240

tokens; two epochs over 4,452 questions. Training feeds each clip to ms-swift as video input: qwen-vl-utils samples it at 2 frames per second, at most 12 frames of at most 448×448 pixels, encoded as video tokens (1,404 video tokens for the first training clip of the seed runs). Evaluation instead sends 12 uniformly sampled JPEG frames of width 448 as separate images, as the benchmark's code does. The reference was trained on one RTX 4080 Super (about 2.2 hours per epoch) with an earlier ms-swift release; the three seeds and the text-only adapter were trained with the same hyperparameters on one 48 GB L40S with ms-swift 4.4.2, transformers 5.12.1, torch 2.14.0, peft 0.19.1 and qwen-vl-utils 0.0.14, using a cosine learning-rate schedule (as logged for those runs). The text-only adapter uses the same training questions without the video. Evaluation: 12 frames uniformly sampled over the clip, resized to width 448, JPEG quality 85, maximum 200 new tokens, temperature 0. The 32B model used 4-bit QLoRA on two 48 GB GPUs with the same data and schedule.