LESS DATA, FASTER CONVERGENCE: GOAL-DRIVEN DATA OPTIMIZATION FOR MULTIMODAL INSTRUCTION TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal instruction tuning can be compute-inefficient when a fixed training budget is spread across large image–video pools whose examples vary substantially in utility. We present Goal-Driven Data Optimization (GDO), a framework that computes six sample descriptors for each candidate and constructs optimized $1\times$ training subsets for different goals. Under a fixed one-epoch Qwen3-VL-8B-Instruct training and evaluation protocol on 32 H20 GPUs, GDO reaches the benchmark-specific UNI-10X references with fewer training examples and attains higher endpoint accuracy. Each reference is the lowest-scoring endpoint among four historical 512k-sample uniform branches, making the comparison descriptive rather than a matched causal estimate. GDO reaches these references after 35.4k samples on MVBENCH, 26.6k on VIDEOMME, 27.3k on MLVU, and 34.7k on LVBENCH, while improving endpoint accuracy by 1.38, 1.67, 3.08, and 0.84 percentage points, respectively. The gains are largest on MVBENCH and MLVU and smaller on LVBENCH, consistent with weaker alignment between its ultra-long-video setting and the short-video and image-dominant training pool. The four profiles span low-loss, coverage-oriented, and temporally focused allocations; among them, Temp+ gives the strongest overall endpoint profile.

1 INTRODUCTION

Modern multimodal assistants have moved from broad visual–language representation learning to instruction-following interaction. Systems such as CLIP Radford et al. (2021), Flamingo Alayrac et al. (2022), BLIP-2 Li et al. (2023a), PaLI Chen et al. (2022), KOSMOS-1 Huang et al. (2023), mPLUG-Owl Ye et al. (2023), LLaVA Liu et al. (2023b), InstructBLIP Dai et al. (2023), MiniGPT-4 Zhu et al. (2023), Qwen-VL Bai et al. (2023), Qwen2-VL Wang et al. (2024), InternVL Chen et al. (2024c), and GPT-4V OpenAI (2023b) have established instruction following as a standard multimodal modeling paradigm. With capable backbones now widely available, a practical question is how to allocate a fixed supervision budget among examples with unequal training value.

This question is sharper in video-centric post-training. Video-capable families such as VideoChat-GPT Maaz et al. (2023), Video-LLaVA Lin et al. (2023), LLaVA-Video Zhang et al. (2024b), LongVA Zhang et al. (2024a), MovieChat Song et al. (2024), and LongViTU Wu et al. (2025), together with large corpora such as Panda-70M Chen et al. (2024a), VISTA Ren et al. (2024), and FineVideo Hugging Face FineVideo Team (2024), expand the available video–language supervision. Yet evaluations on MVBENCH Li et al. (2024d), VIDEOMME Fu et al. (2025), MLVU Zhou et al. (2025), and LVBENCH Wang et al. (2025) continue to reveal uneven performance across motion, ordering, and long-context reasoning tasks. Related work on open-world visual reasoning likewise shows that free-form concept discrimination remains challenging Wu et al. (2023).

This creates a post-training allocation problem. Mixed image–video instruction pools are large, redundant, and heterogeneous, and the value of an additional training example is not uniform. Some samples repeat already-solved patterns; others supply temporal evidence, ordering pressure, or reliable supervision that changes downstream behavior. Prior work on instruction tuning and alignment also shows that curation quality, diversity, and task alignment can matter as much as dataset size in the evaluated settings Taori et al. (2023); Zhou et al. (2023); Chen et al. (2023a); Li et al. (2023b).

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.

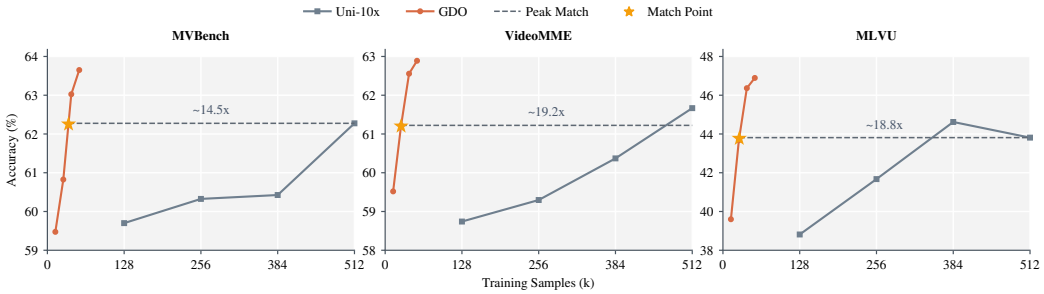


Figure 1: **Peak Match with Less Data.** Accuracy is plotted against training samples for MVBENCH, VIDEOMME, and MLVU. The dashed line marks the fixed UNI-10X reference, and the star marks its first crossing under linear interpolation between adjacent displayed GDO checkpoints. Table 1 reports the corresponding Peak Match and Reduction values.

Scaling-law work Kaplan et al. (2020); Hoffmann et al. (2022); Sorscher et al. (2022), data-quality studies Wenzek et al. (2020); Abbas et al. (2023); Li et al. (2024c); Liu et al. (2023c), and instruction-data selection pipelines Xia et al. (2024); Kung et al. (2023) reinforce this point. We ask whether goal-driven data optimization can reduce the supervision needed for multimodal instruction tuning while accelerating convergence under a fixed training protocol.

Existing methods address complementary aspects of data efficiency, but they do not by themselves provide the fixed comparison used in this study. Curriculum and pacing methods Bengio et al. (2009); Kumar et al. (2010); Graves et al. (2017); Fan et al. (2018), sample reweighting Jiang et al. (2018); Ren et al. (2018), active selection Settles (2009); Sener & Savarese (2018); Kirsch et al. (2019), valuation Ghorbani & Zou (2019); Jia et al. (2019); Kwon & Zou (2023), and instruction-data selection pipelines Moore & Lewis (2010); Xia et al. (2024); Kung et al. (2023) all show that data choice matters. We instead isolate data allocation from changes in model, optimizer, and evaluation protocol.

We answer this question with Goal-Driven Data Optimization (GDO). GDO changes only the selected training subset while holding the model, optimizer, checkpoint schedule, and evaluation protocol fixed. It computes six descriptors for each candidate and constructs optimized $1\times$ subsets subject to explicit budget, modality-mixture, and source-coverage constraints. This design makes observed performance shifts interpretable as data-allocation effects rather than changes in the surrounding recipe.

The four released goal profiles share one descriptor extraction and scoring backbone, and differ primarily in subset size, video-ratio bands, VDS-positive coverage targets, source floors, and over-sampling schedules. This design provides a controlled framework for testing how data-optimization goals affect capability, convergence, and sample efficiency under the same training contract.

Two empirical asymmetries motivate this design. First, quantity and usefulness are decoupled in mixed image–video instruction pools. Increasing short-video and image QA by count alone can improve already-strong perception categories while leaving temporal reasoning, ordering, and evidence integration under-served. Second, benchmark alignment is uneven. MVBENCH and MLVU organize evaluation around focused subtasks involving motion, ordering, and comparison. By contrast, LVBENCH emphasizes ultra-long-video understanding and is less closely matched to the short-video and image-dominant training pool used here. This mismatch offers one plausible explanation for the smaller gains observed on LVBENCH. Thus, GDO is a controlled data-allocation framework whose effect depends on the match between the candidate pool and target evaluation distribution.

Sample efficiency and convergence must be evaluated together. A small optimized subset is meaningful only if it reaches useful capability earlier without trading final quality for speed. Conversely, a final-score gain is incomplete if it requires a larger training budget. Our evaluation therefore freezes one UNI-10X reference per benchmark and asks whether goal-driven data optimization can reach or exceed that reference earlier while attaining stronger reported endpoints.

Table 1: **Peak Match and Reduction.** Peak Match is the descriptive first crossing of the fixed UNI-10X reference under linear interpolation between adjacent displayed checkpoints; it is not an independently evaluated checkpoint. Reduction is relative to the 512k-sample UNI-10X budget. Scores are accuracy (%), and Δ is the improvement over UNI-10X in percentage points (pp).

Benchmark	UNI-10X	GDO	Δ (pp)	Peak Match	Reduction
MVBENCH	62.27	63.65	*+1.38*	35.4k	14.5×
VIDEOMME	61.22	62.89	*+1.67*	26.6k	19.2×
MLVU	43.81	46.89	*+3.08*	27.3k	18.8×
LVBENCH	40.22	41.06	*+0.84*	34.7k	14.8×

Figure 1 shows that a smaller optimized subset can reach and exceed the fixed UNI-10X reference earlier in training. Table 1 reports more than 10× peak-match data reduction on all four benchmarks. The gains vary across benchmarks: they are largest on MVBENCH and MLVU and smaller on LVBENCH. This pattern is consistent with closer alignment between the available pool and the focused subtasks in MVBENCH and MLVU, and weaker alignment with the ultra-long-video regime of LVBENCH.

Taken together, the results support three claims:

- **Less Data**. Optimized 1× subsets can outperform the fixed UNI-10X baseline with far fewer training samples.
- **Faster Convergence**. The gains appear as earlier frontier crossings, not only as stronger reported endpoints.
- **Goal-Driven Data Optimization**. Different allocation goals produce distinct capability profiles, with Temp+ giving the strongest overall result among the tested profiles.

Our contributions are therefore threefold:

- We formulate multimodal instruction-data allocation under fixed training and evaluation settings as a goal-driven data optimization problem.
- We instantiate GDO as a reproducible pipeline that combines six sample descriptors, shared candidate scoring, and goal-specific optimization presets over budget, video composition, coverage, and source balancing.
- We provide benchmark, trajectory, subtask, and ablation evidence that the selected subsets reach the fixed reference earlier and attain higher endpoint accuracy under this contract, while documenting how the gains vary across benchmarks.

2 METHOD

2.1 PROBLEM SETUP

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ denote the shared multimodal instruction pool, g a data-allocation goal, and b an evaluation benchmark. For each goal, GDO builds a 1× optimized subset S_g. For each benchmark, U_b denotes the frozen endpoint score of the weakest 512k-sample uniform branch among the four strategy suites. Thus U_b is a benchmark-specific reporting reference, not one shared uniform run. The backbone, one-epoch SFT recipe, and evaluation implementation remain fixed across all branches.

The released pipeline is deterministic and staged. It first computes six sample descriptors, then maps each candidate to a shared score $\rho(x)$, and finally applies a goal-specific feasibility preset $\mathcal{C}_g$ that controls budget, video ratio, temporal-positive coverage, source floors, and candidate oversampling as follows:

$$S_g = \mathrm{Build}(\mathcal{D}; \rho, \mathcal{C}_g)\,, \qquad \Delta_{g,b} = \mathrm{Eval}_b(S_g) - U_b. \tag{1}$$

At the paper level, the intervention is restricted to subset construction. As summarized in Figure 2, the descriptors provide comparable cues, the shared score defines preference, and the goal preset defines admissibility. Subset construction is benchmark-blind: benchmark test identities and answers never enter descriptor extraction, scoring, or filtering. The same descriptor table is reused across

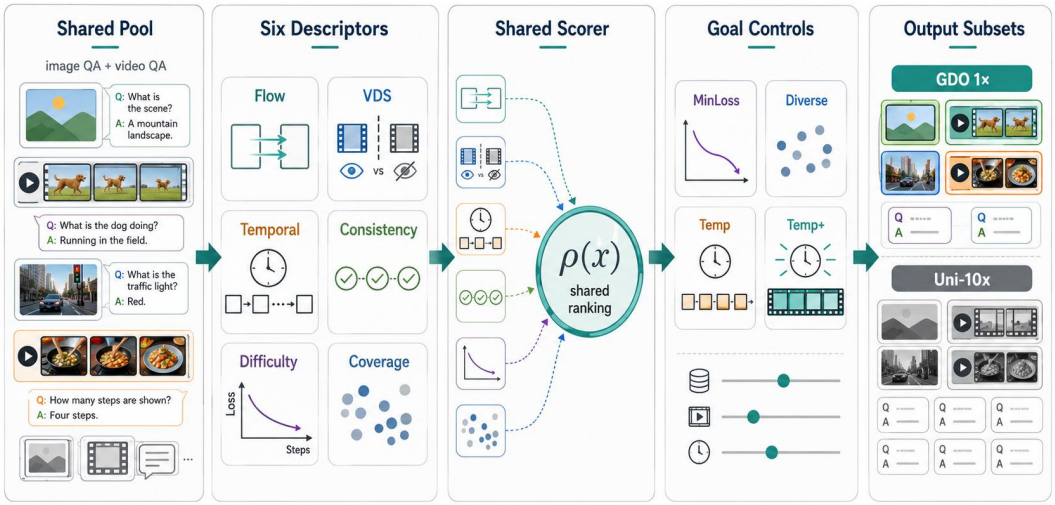


Figure 2: **Subset Construction.** GDO computes six sample descriptors over one shared pool, applies a shared score and goal-specific feasibility presets, and trains the resulting $1\times$ subsets under the same backbone and SFT recipe. Evaluation uses the frozen benchmark-specific UNI-10X references defined in Section 3.

profiles. Benchmark-wise selection of U_b affects the reporting reference, not subset construction, and is disclosed throughout.

2.2 SIX SAMPLE DESCRIPTORS

Each candidate is represented by a six-dimensional descriptor vector

$$\mathbf{m}(x) = [m_{\text{flow}}, m_{\text{vds}}, m_{\text{tnc}}, m_{\text{sc}}, m_{\text{ppl}}, m_{\text{cov}}]\,. \tag{2}$$

All six descriptors are computed from candidate training examples with frozen probes or deterministic statistics; none uses benchmark labels, test examples, or downstream rewards.

For a sample $x = (v_1, \ldots, v_T, q, y)$, GDO instantiates the six descriptors as follows:

Flow measures average optical-flow magnitude across adjacent frames

$$m_{\text{flow}}(x) = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{1}{|\Omega|}\sum_{p\in\Omega}\|u_t(p)\|_2, \tag{3}$$

where u_t is the dense flow field between frames v_t and v_{t+1}. We average per-pixel motion magnitude over adjacent frame pairs; larger values therefore correspond to stronger visible motion cues.

VDS (video-dependence score) uses the blind-vs-video loss gap

$$m_{\text{vds}}(x) = \mathcal{L}_{\text{blind}}(x) - \mathcal{L}_{\text{video}}(x), \tag{4}$$

where $\mathcal{L}_{\text{blind}}$ and $\mathcal{L}_{\text{video}}$ are computed with the same frozen Qwen3-VL-8B-Instruct probe under blind and video-conditioned inputs. We evaluate each question–answer pair with and without video frames; a larger VDS means that visual input reduces the answer loss more strongly.

Temporal necessity assigns a question-level proxy score

$$m_{\text{tnc}}(x) = T(q), \tag{5}$$

where $T(q) \in [0, 1]$ is a temporal-necessity judgment from the frozen probe. We map its judgment to a scalar proxy; higher values indicate that the question is more likely to require ordering, change, duration, or cross-frame reasoning rather than static visual lookup.

Self-consistency measures agreement across stochastic decodes

$$m_{\text{sc}}(x) = \text{mean}_{i<j}\ \text{Jac}(a_i, a_j), \tag{6}$$

where a_i are stochastic decodes. We compute mean pairwise Jaccard similarity across repeated sampled answers; higher agreement indicates a more stable probe response.

PPL-like difficulty uses the exponentiated teacher-forced video loss

$$m_{\mathrm{ppl}}(x) = \exp(\mathcal{L}_{\mathrm{video}}(x)). \tag{7}$$

We exponentiate the teacher-forced video-conditioned loss of the target answer; larger values indicate that the frozen probe finds the sample harder to fit.

Coverage combines semantic clustering and source statistics

$$m_{\mathrm{cov}}(x) = C(x; \mathcal{N}_{\mathrm{text}}, \mathcal{N}_{\mathrm{vision}}, s(x)), \tag{8}$$

where $C(\cdot)$ combines the local text neighborhood $\mathcal{N}_{\mathrm{text}}$, visual neighborhood $\mathcal{N}_{\mathrm{vision}}$, and source identity $s(x)$. Larger values favor semantic and source-level breadth over already over-represented regions of the pool.

The builder consumes these descriptors in two complementary forms, with their normalized combinations contributing to a merged quality summary used by the shared scorer while the raw descriptors remain available to feasibility controls and ablations. As a result, one descriptor layer supports both global ranking and goal-specific composition constraints, while still covering motion content, video dependence, temporal demand, answer stability, optimization difficulty, and semantic/source breadth. GDO therefore preserves multiple allocation cues that would otherwise be compressed into a single quality score.

2.3 Scoring and Feasibility

The descriptor table is used in two coupled but distinct ways. A shared score $\rho(x)$ determines the candidate ranking, while the goal-specific preset $\mathcal{C}_g$ determines which ranked candidates are admissible under a given allocation target. The four profiles share one scoring backbone and differ mainly in budget and composition constraints. For short-video samples, the released score is

$$\rho_{\mathrm{vid}}(x) = 0.35 \tanh\left(\frac{b_{\mathrm{vid}}(x)}{3}\right) + 0.95\, z_{\mathrm{vds3}}(x) + 0.35\, z_{\mathrm{qual}}(x), \tag{9}$$

where z_{vds3} is a normalized video-dependence term derived from $\mathcal{L}_{\mathrm{video}}$, $\mathcal{L}_{\mathrm{blind}}$, and frame diversity, and z_{qual} is the normalized merged `quality_score`. The base term is

$$b_{\mathrm{vid}}(x) = q_{\mathrm{text}}(x) + 0.85\, d(x) + 0.9\, a(x) + 0.55\, t(x) + 0.15\, r_{\mathrm{src}}(x) \tag{10}$$

where q_{text} is a heuristic text-quality prior, d is a medium-difficulty preference, a is a bin-alignment term over duration, temporal, question-length, and source strata, t is a temporal bonus, and r_{src} is a source-rarity term.

For image QA, the released builder uses the same logic without the video-only terms

$$\rho_{\mathrm{img}}(x) = 0.90 \tanh\left(\frac{b_{\mathrm{img}}(x)}{3}\right) + 0.15\, z_{\mathrm{qual}}(x), \tag{11}$$

with $b_{\mathrm{img}}(x)$ omitting the temporal bonus and using a slightly stronger text-quality weight. Because the contributing terms are normalized before mixing, these coefficients should be read as fixed mixture weights in the released builder rather than as learned parameters. They are set once to balance the base heuristic, video-dependence, and merged-quality contributions, and are then shared across all four goal profiles rather than retuned per goal or benchmark. The ablations later in the paper therefore test component importance under one fixed scorer instead of re-optimizing these weights.

Because ρ is shared, changing g primarily changes admissibility. A global sort by $\rho(x)$ would over-select abundant, linguistically clean image QA and under-select rarer temporal video samples. The feasibility controls instead enforce modality mix, temporal coverage, and source breadth alongside score. The same descriptor table and scorer can therefore support multiple allocation goals by changing only the feasibility constraints.

Operationally, the builder maintains per-stratum top-k reservoirs and fills the subset in stages that include temporal-video minima, video-ratio minima, source floors, stratum quotas, and finally the global score tail with reservoir fallback. Deduplication and QA-per-video caps are enforced throughout. This staged fill is the released realization of Eq. (1).

Table 2: **Overall Results.** Rows report the four released 1× GDO profiles. Each cell gives accuracy (%), with the pp difference relative to the frozen benchmark-specific 512k-sample UNI-10X reference in parentheses.

Setting	1× samples	MVBench	VideoMME	MLVU	LVBench
MinLoss	12.9k	63.63 (*+1.35*)	62.30 (*+1.07*)	45.84 (*+2.03*)	38.86 (*-1.36*)
Diverse	42.9k	63.12 (*+0.85*)	61.33 (*+0.11*)	46.05 (*+2.24*)	39.90 (*-0.32*)
Temp	33.3k	62.05 (*-0.23*)	62.04 (*+0.81*)	45.26 (*+1.45*)	40.28 (*+0.06*)
Temp+	53.3k	**63.65** (*+1.38*)	**62.89** (*+1.67*)	**46.89** (*+3.08*)	**41.06** (*+0.84*)

2.4 GOAL PROFILES

The paper reports four released goal profiles within GDO, each corresponding to a different setting of the same builder with shared scoring and different feasibility presets. For paper reporting, all four profiles are compared against the frozen benchmark-specific UNI-10X references:

MinLoss explicitly targets the lowest training loss during subset construction. It therefore favors the easiest-to-fit supervision under the smallest budget (12.9k) and applies the lightest temporal and video pressure.

Diverse shifts the objective from minimum loss to broader semantic and source coverage. Its larger subset (42.9k) and stronger source floors make it the most conservative profile against collapse onto only the easiest samples.

Temp optimizes for stronger temporal usefulness. It allocates 33.3k samples while raising the target video ratio and temporal-positive coverage so that more of the selected budget is spent on temporally informative video supervision.

Temp+ pushes that temporal objective further. It increases both subset size and temporal pressure (53.3k, 58.7% selected video), making it the strongest temporal profile in the released suite.

Together, the profiles span low-loss, coverage-oriented, and temporally focused allocations. They share ρ and differ through the controls in $\mathcal{C}_g$, allowing the resulting frontier shifts to be compared without changing the scorer.

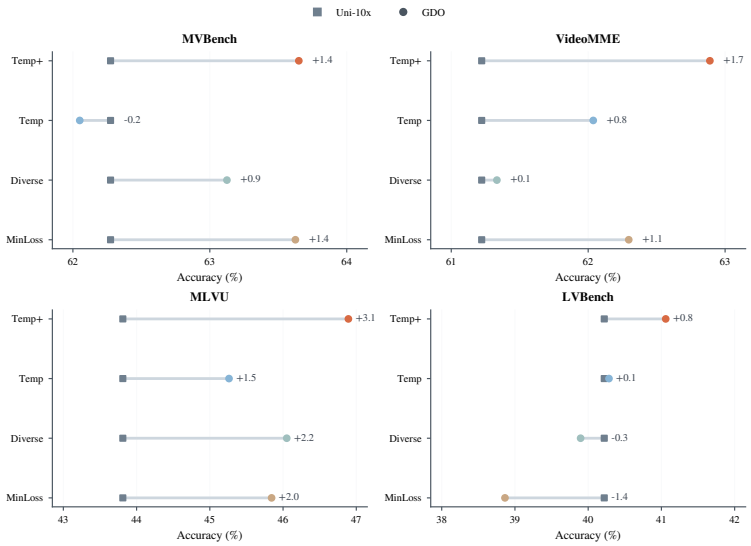


Figure 3: **Frontier Shifts by Goal.** For each benchmark, dots mark the strongest operating points of the four GDO profiles, while the square marks the frozen benchmark-specific UNI-10X reference. These are descriptive comparisons, not matched causal contrasts.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Data. We build all subsets from one shared pool consisting of the full LLaVA-OneVision image-QA data Li et al. (2024a) and the full LLaVA-Video short-video QA data Zhang et al. (2024b). Each GDO profile constructs one optimized $1\times$ subset from that pool.

Training. We follow one fixed one-epoch Qwen3-VL-8B-Instruct SFT Bai et al. (2025) recipe on 32 H20 GPUs with batch size 1 per GPU and global batch size 32. Only the selected subset and its resulting sample count vary across GDO profiles.

Comparison and trajectories. For each benchmark, we freeze the lowest-scoring endpoint among four historical 512k-sample UNI-10X branches as U_b and use it for every reported delta. Consequently, the comparisons are descriptive and are not matched causal controls. For trajectory figures, four points are retained from five saved checkpoints using a fixed screening rule that removes one noisy point and preserves the most monotone subsequence. Their measured scores are unchanged, but their horizontal positions are pseudo-quarter sample counts based on each branch's maximum paper-facing budget. Endpoint scores are direct checkpoint measurements; intermediate positions and peak-match crossings are descriptive visual summaries.

Benchmarks. We evaluate MVBENCH Li et al. (2024d), VIDEOMME Fu et al. (2025), MLVU Zhou et al. (2025), and LVBENCH Wang et al. (2025). MVBENCH and MLVU are subtask-oriented, VIDEOMME is broader and less tightly factorized, and LVBENCH emphasizes ultra-long-video understanding. This combination tests whether the same data-allocation strategy transfers across focused temporal subtasks, broader video QA, and a substantially longer-video regime. All scores are reported as accuracy (%), with differences in percentage points (pp).

3.2 BENCHMARK RESULTS

Table 2 gives the main benchmark result. With fewer than 512k training samples, GDO improves MVBENCH, VIDEOMME, MLVU, and LVBENCH by 1.38, 1.67, 3.08, and 0.84 pp, respectively. The largest gains occur on MVBENCH and MLVU, whose focused subtasks emphasize motion, order, and temporal evidence. This association is consistent with the intended allocation, although the experiment does not isolate benchmark structure as its cause. VIDEOMME also improves, showing that the gains are not confined to one subtask-heavy evaluation style.

Figure 3 shows the same result from the convergence side. Linear interpolation between adjacent displayed checkpoints places the first reference crossing at 35.4k samples on MVBENCH, 26.6k on VIDEOMME, 27.3k on MLVU, and 34.7k on LVBENCH. These descriptive peak-match points operationalize "less data, faster convergence" at the recorded checkpoint cadence; they are not independently evaluated checkpoints. Together, Table 2 and Figure 3 show changes in both endpoint accuracy and the observed training trajectory.

Looking across the four profiles is more informative than focusing only on the headline GDO result. MinLoss targets the lowest-loss supervision and therefore reaches useful operating points under the smallest budget. It isolates what low-loss allocation alone provides, namely strong sample efficiency with limited emphasis on the harder temporal evidence required by the most demanding temporal tasks. Diverse behaves differently. It restores broader semantic and source coverage, which keeps it close to the fixed UNI-10X reference on several benchmarks and avoids collapsing onto only the easiest supervision, but it remains conservative on temporal specialization.

Temp and Temp+ allocate more of the budget to temporally informative video supervision. The intermediate profiles are not strictly ordered on every benchmark, but Temp+ gives the strongest endpoint result on all four. The profile comparison therefore supports a difference among allocation goals without implying a monotonic response at every intermediate setting.

The LVBENCH result clarifies the scope of the method. It evaluates ultra-long videos, often much longer than the clips represented in LLaVA-Video, while the training pool also contains substantial image QA from LLaVA-OneVision. Within the tested profiles, LVBENCH changes from negative or near-zero differences for MinLoss and Diverse to positive differences for Temp and Temp+. This

Table 3: **Subtask Deltas.** Values are subtask deltas in pp relative to the frozen benchmark-specific UNI-10X reference. Unlike Table 4, which highlights the largest temporal wins, this table keeps adverse subtasks visible.

Setting	MVB Motion	MVB Reasoning	VMM Temp. Perc.	MLVU Order	MLVU SportsQA	MLVU Ego
MinLoss	*-1.1*	*+3.2*	*+9.1*	*+5.7*	*+11.1*	*-15.1*
Diverse	*+1.5*	*-2.2*	*+10.9*	*+10.0*	*+8.3*	*-5.7*
Temp	*-1.2*	*+0.6*	*+7.3*	*+4.3*	*+5.6*	***+1.9***
Temp+	***+0.8***	***+3.8***	***+18.2***	***+7.1***	***+11.1***	*-3.8*

pattern is consistent with a benefit from temporal allocation, while the available pool limits the size of the observed gain.

3.3 SUBTASK ANALYSIS

Table 3 then shows the same movement over the main subtasks while keeping adverse subtasks visible. Temp+ is strongest because it combines positive motion, positive reasoning, and very large temporal-perception gains without giving up the strongest order-related subtasks. Diverse preserves some motion and order behavior while remaining more coverage-oriented, whereas MinLoss trades isolated wins for larger regressions elsewhere. Temp sits between them and makes the transition from broad data allocation to explicitly temporal allocation visible.

3.4 ABLATION STUDY

The profile comparison already shows that stronger temporal goals matter. We next ask whether Temp+ is driven by one dominant score term or by the combination of several cues. Table 5 supports the latter reading.

The ablation pattern is benchmark-dependent rather than collapsing onto one dominant feature. On VIDEOMME, removing VDS or the PPL-like term costs about one point. On MVBENCH and especially MLVU, removing self-consistency produces the largest decrease among the single-component ablations, which is consistent with answer stability being a useful cue in this setting. Some single removals help one benchmark while hurting others, further arguing against a single monotone driver. The joint removal is more damaging on MVBENCH than any individual ablation, suggesting that Temp+ relies on multiple scoring ingredients and feasibility controls.

3.5 DISCUSSION

The optimized subsets change both the endpoint and the observed training trajectory: they reach the fixed UNI-10X reference at earlier checkpoints and finish above it under the same backbone, optimizer, and evaluation protocol.

The gains are capability-dependent. MVBENCH and MLVU benefit most, while LVBENCH improves more modestly under the short-video and image-heavy training pool. The result is therefore conditional on the candidate pool: allocation can only emphasize capabilities for which the pool contains relevant supervision.

The released profiles form an allocation spectrum rather than a single operating point. MinLoss is the most budget-efficient setting, Diverse restores broader coverage, and Temp/Temp+ place greater emphasis on temporal supervision. The subtask and ablation results show concentrated gains alongside adverse cells. Under the fixed protocol studied here, data allocation changes the efficiency–capability trade-off, and the appropriate profile depends on the target capability and source distribution.

4 Related Work

CLIP and CoCa establish large-scale image–text pretraining Radford et al. (2021); Yu et al. (2022). Flamingo, BLIP-2, PaLI, PaLI-X, KOSMOS-1, and mPLUG-Owl extend this line toward few-shot behavior, generation, and unified multimodal modeling Alayrac et al. (2022); Li et al. (2023a); Chen et al. (2022; 2023b); Huang et al. (2023); Ye et al. (2023). LLaVA, LLaVA-1.5, LLaVA-NeXT-Interleave, InstructBLIP, and MiniGPT-4 turn such backbones into assistants Liu et al. (2023b;a); Li et al. (2024b); Dai et al. (2023); Zhu et al. (2023), while Qwen-VL, Qwen2-VL, InternVL, InternVL 2.5, GPT-4, and GPT-4V further strengthen this family Bai et al. (2023); Wang et al. (2024); Chen et al. (2024c;b); OpenAI (2023a;b). These systems motivate strong fixed backbones, but leave open how to allocate a limited multimodal SFT budget once the backbone is fixed.

Alpaca, LIMA, and AlpaGasus study compact or curated instruction sets Taori et al. (2023); Zhou et al. (2023); Chen et al. (2023a). M3IT and InstructBLIP extend instruction tuning to heterogeneous multimodal sources Li et al. (2023b); Dai et al. (2023). GDO combines this motivation with benchmark-specific frozen reporting references; the resulting contrasts are descriptive rather than matched causal estimates.

VideoChatGPT, Video-LLaVA, Chat-UniVi, MovieChat, and LLaVA-Video extend instruction following from images to videos Maaz et al. (2023); Lin et al. (2023); Jin et al. (2024); Song et al. (2024); Zhang et al. (2024b). LongVA, VideoAgent, and LongViTU push toward longer context Zhang et al. (2024a); Fan et al. (2024); Wu et al. (2025), while Panda-70M, VISTA, and FineVideo enlarge the video–language supervision pool Chen et al. (2024a); Ren et al. (2024); Hugging Face FineVideo Team (2024). MVBench, VideoMME, MLVU, and LVBench expose uneven performance across temporal subtasks Li et al. (2024d); Fu et al. (2025); Zhou et al. (2025); Wang et al. (2025), making data allocation relevant to video-centric post-training.

Scaling analyses study compute–data balance Kaplan et al. (2020); Hoffmann et al. (2022); Sorscher et al. (2022). Curriculum and pacing methods change when examples are seen Bengio et al. (2009); Kumar et al. (2010); Graves et al. (2017); Fan et al. (2018), while sample reweighting and forgetting analyses study which examples matter most Jiang et al. (2018); Ren et al. (2018); Toneva et al. (2018); Paul et al. (2021). Active selection and data valuation make that choice explicit Settles (2009); Sener & Savarese (2018); Kirsch et al. (2019); Ghorbani & Zou (2019); Jia et al. (2019); Kwon & Zou (2023). Corpus filtering and pre-training data management improve large training pools Wenzek et al. (2020); Abbas et al. (2023); Li et al. (2024c); Peng et al. (2025); Zhuang et al. (2025), while instruction-data selection and refinement target post-training supervision Liu et al. (2023c); Moore & Lewis (2010); Xia et al. (2024); Kung et al. (2023); Zhao et al. (2024); Lyu et al. (2025); Pang et al. (2025). These studies motivate data-quality and selection signals. GDO combines multiple descriptors with goal-specific feasibility constraints. Self-consistency provides a complementary stability cue Wang et al. (2022).

5 Conclusion

We presented GDO, a goal-driven data optimization framework for multimodal instruction tuning. Under the fixed Qwen3-VL-8B-Instruct protocol, GDO reaches the frozen benchmark-specific Uni-10x references with fewer samples and delivers higher reported endpoint accuracy on MVBench, VideoMME, MLVU, and LVBench. Because each reference is the weakest historical uniform branch for that benchmark, these comparisons describe the original reporting protocol rather than matched causal effects. The released profiles show distinct efficiency–capability trade-offs.

These results frame multimodal SFT data as an allocation problem as well as a scaling problem. Combining sample descriptors with shared scoring and goal-specific feasibility controls exposes that choice while keeping the remaining training protocol fixed. The conclusion is bounded by the candidate pool: improvements are largest when its supervision is aligned with the target capability, whereas ultra-long-video evaluation remains weakly matched by the data used here.

Reproducibility Statement

The main text specifies the candidate pool, subset sizes, backbone, training protocol, checkpoints, evaluation benchmarks, and reported metrics. The appendix provides the shared scoring functions, goal-profile constraints, full trajectories, subtask analyses, ablations, limitations, and qualitative examples. An anonymized code package contains the data-processing, subset-construction, training, evaluation, and figure-generation implementations.

AI Use Statement

Generative AI tools were used primarily to check grammar and improve the clarity and readability of the manuscript. The authors reviewed all AI-assisted edits and take responsibility for the final manuscript.

References

Amro Abbas, Kushal Tirumala, Daniel Simig, et al. Semdedup: Data-efficient learning at web scale through semantic deduplication. In *International Conference on Learning Representations*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: A visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yoshua Bengio, Jerome Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009.

Lichang Chen, Shiyang Li, Jun Yan, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023a.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024a.

Xi Chen, Xiao Wang, Soravit Changpinyo, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Xi Chen, Josip Djolonga, Piotr Padlewski, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023b.

Zhe Chen, Weiyun Wang, Yue Cao, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2024c.

Wenliang Dai, Junnan Li, Dongxu Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

Yitong Fan, Fei Tian, Tao Qin, et al. Learning to teach. In *International Conference on Learning Representations*, 2018.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *arXiv preprint arXiv:2403.11481*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. doi: 10.1109/CVPR52734.2025.02245.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, 2019.

Alex Graves, Marc G. Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. *arXiv preprint arXiv:1704.03003*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Shaohan Huang, Li Dong, Wenhui Wang, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

Hugging Face FineVideo Team. Finevideo: A dataset with rich descriptions, narrative details, scene splits, and qa pairs. Hugging Face dataset and project page, 2024. URL `https://github.com/huggingface/fineVideo`. Accessed: 2026-06-30.

Ruoxi Jia, David Dao, Boxin Wang, et al. Towards efficient data valuation based on the shapley value. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, 2018.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, et al. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, 2019.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, 2010.

Po-Nien Kung, Fan Yin, Di Wu, et al. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. *arXiv preprint arXiv:2311.00288*, 2023.

Yongchan Kwon and James Zou. Data-oob: Out-of-bag estimate as a simple and efficient data value. *arXiv preprint arXiv:2304.07718*, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Feng Li, Renrui Zhang, Hao Zhang, et al. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Jeffrey Li, Alex Fang, Georgios Smyrnis, et al. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024c.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Kunchang Li, Yali Wang, Yinan He, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024d. doi: 10.1109/CVPR52733.2024.02095.

Lei Li, Yuwei Yin, Shicheng Li, et al. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023b.

Bin Lin, Yang Ye, Bin Zhu, et al. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Wei Liu, Weihao Zeng, Keqing He, et al. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*, 2023c.

Mengyao Lyu, Yan Li, Huasong Zhong, Wenhao Yang, Hui Chen, Jungong Han, Guiguang Ding, and Zhenheng Yang. Cream of the crop: Harvesting rich, scalable and transferable multi-modal data for instruction fine-tuning. *arXiv preprint arXiv:2503.13383*, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *ACL Short Papers*, 2010.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023a.

OpenAI. Gpt-4v system card. *Technical report*, 2023b.

Jinlong Pang, Jiaheng Wei, Ankit Parag Shah, Zhaowei Zhu, Yaxuan Wang, Chen Qian, Yang Liu, Yujia Bao, and Wei Wei. Improving data efficiency via curating LLM-driven rating systems. In *International Conference on Learning Representations (ICLR)*, 2025.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *arXiv preprint arXiv:2107.07075*, 2021.

Ru Peng, Kexin Yang, Yawen Zeng, Junyang Lin, Dayiheng Liu, and Junbo Zhao. Dataman: Data manager for pre-training large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, 2018.

Weiming Ren, Huan Yang, Jie Min, et al. Vista: Enhancing long-duration and high-resolution video understanding by video spatiotemporal augmentation. *arXiv preprint arXiv:2412.00927*, 2024.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Burr Settles. Active learning literature survey. *Computer Sciences Technical Report 1648, University of Wisconsin–Madison*, 2009.

Enxin Song, Wenhao Chai, Guanhong Wang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2024.

Ben Sorscher, Robert Geirhos, Sneha Shekhar, et al. Beyond neural scaling laws: Beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 2022.

Rishi Taori, Ishaan Gulrajani, Tianyi Zhang, et al. Stanford alpaca: An instruction-following llama model. *Technical report*, 2023.

Mariya Toneva, Alessandro Sordoni, Remi Combes, et al. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2018.

Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025. doi: 10.1109/ICCV51701.2025.02131.

Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, et al. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *International Conference on Language Resources and Evaluation*, 2020.

Rujie Wu, Xiaojian Ma, Zhenliang Zhang, Wei Wang, Qing Li, Song-Chun Zhu, and Yizhou Wang. Bongard-openworld: Few-shot reasoning for free-form visual concepts in the real world. *arXiv preprint arXiv:2310.10207*, 2023.

Rujie Wu, Xiaojian Ma, Hai Ci, Yue Fan, Yuxuan Wang, Haozhe Zhao, Qing Li, and Yizhou Wang. Longvitu: Instruction tuning for long-form video understanding. *arXiv preprint arXiv:2501.05037*, 2025.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, et al. Less: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning*, 2024.

Qinghao Ye, Haiyang Xu, Guohai Xu, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, et al. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Peiyuan Zhang, Kaichen Zhang, Bo Li, et al. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, et al. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *arXiv preprint arXiv:2407.05282*, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 2023.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025. doi: 10.1109/CVPR52734.2025.01278.

Deyao Zhu, Jun Chen, Xiaoqian Shen, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Jiantao Qiu, Chi Zhang, Ying Qian, and Conghui He. Meta-rater: A multi-dimensional data selection method for pre-training language models. *arXiv preprint arXiv:2504.14194*, 2025.

APPENDIX

A ADDITIONAL ANALYSIS

A.1 TRAJECTORIES

Figure 4 extends the peak-match result to full trajectories on all four benchmarks under the same fixed comparison. The earlier crossing pattern is visible beyond a single checkpoint. The separation is largest on MVBENCH, VIDEOMME, and MLVU and smaller on LVBENCH. The latter pattern is consistent with, but does not by itself establish, a mismatch between the ultra-long evaluation videos and the short-video and image-dominant training pool.

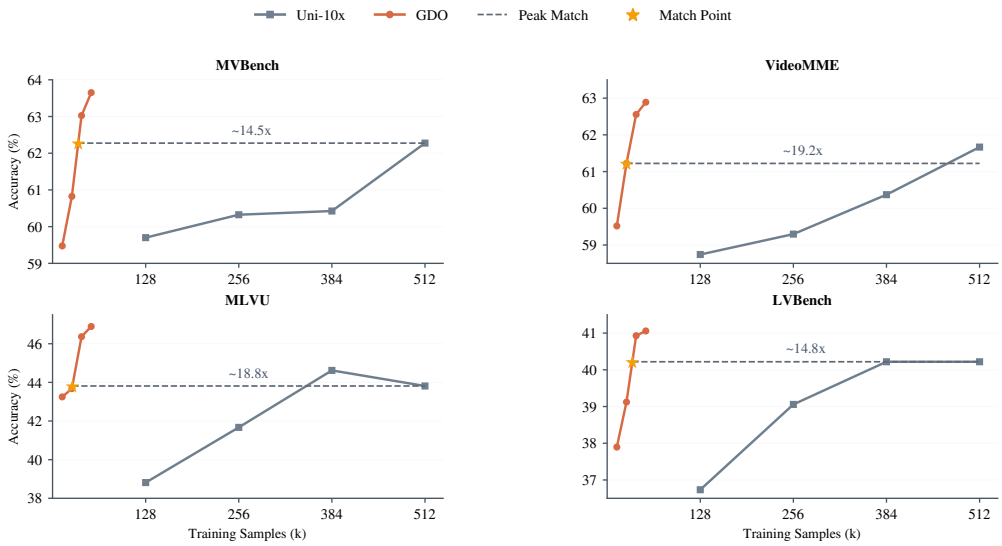


Figure 4: **Trajectories.** Each panel compares GDO with its frozen benchmark-specific 512k-sample UNI-10X reference. Stars denote the first linearly interpolated crossing between adjacent displayed checkpoints.

A.2 GOAL PROFILES

Figure 5 places the four released goal profiles against the same frozen per-benchmark UNI-10X reference. MinLoss occupies the earliest low-budget regime, whereas Diverse stays closer to a broad-coverage allocation. Temp and Temp+ raise both the selected video ratio and the temporal-positive floor; Temp+ attains the strongest endpoint profile, although the intermediate profiles are not strictly ordered on every benchmark.

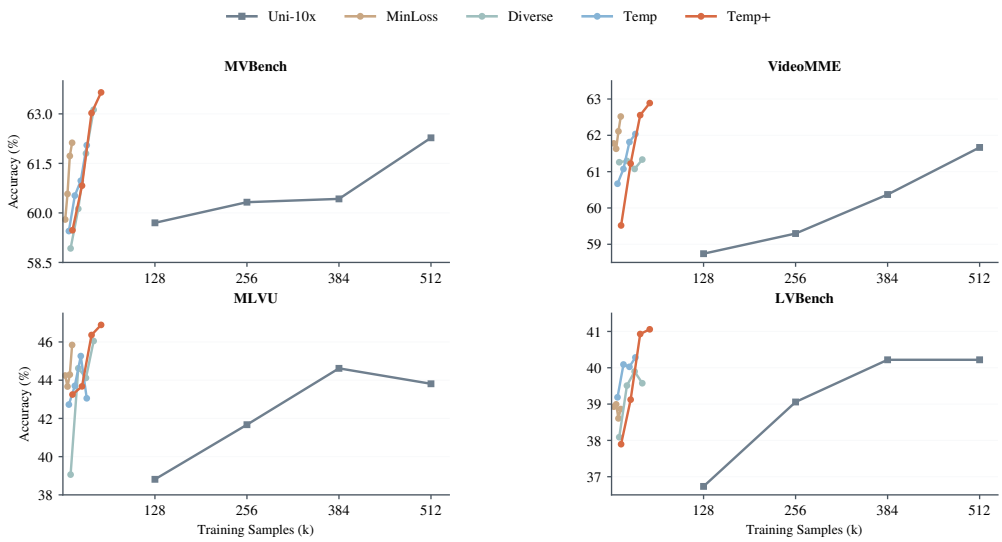


Figure 5: **Goal Profiles.** Each panel overlays MinLoss, Diverse, Temp, and Temp+ against the same frozen benchmark-specific UNI-10X reference.

A.3 ABLATION

The profile comparison does not identify which parts of the Temp+ builder contribute most. Figure 6 extends the main-paper ablation into a trajectory comparison while keeping the remaining training and evaluation protocol fixed. On VIDEOMME, removing VDS or PPL hurts most. On MLVU, removing self-consistency causes the largest drop among the tested single-component removals, which is consistent with answer stability being useful in this setting. On MVBENCH, the joint removal is most damaging, showing that no single ablation accounts for the full change.

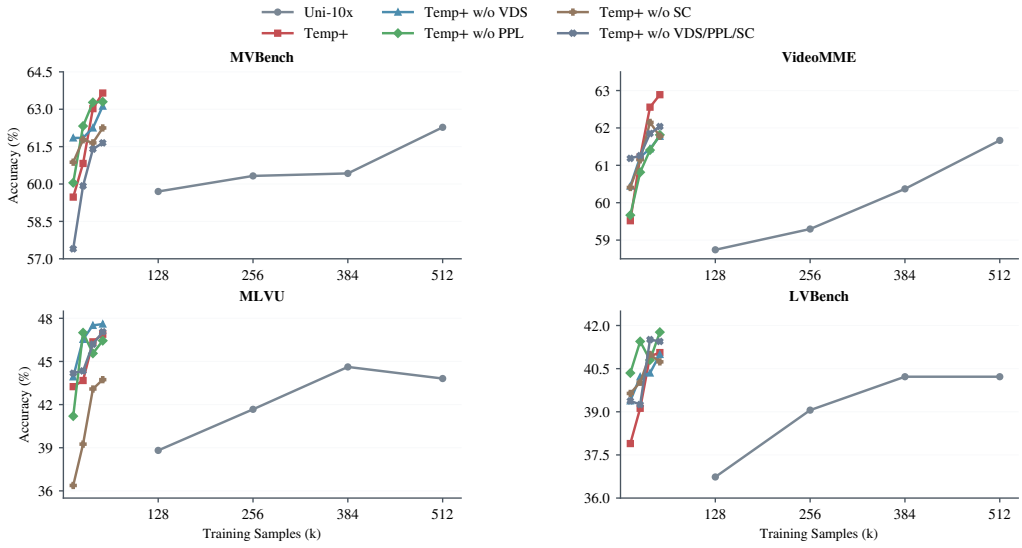


Figure 6: **Ablation Trajectories.** Each curve removes one or more Temp+ score contributions while keeping the rest of the train/eval contract fixed.

Table 4: **Temporal Subtask Gains.** The five temporally related subtasks with the largest gains for each GDO profile over the fixed per-benchmark UNI-10X reference. Scores are accuracy (%), and Δ is the difference in percentage points.

Benchmark	Subtask	UNI-10X (%)	GDO (%)	Δ (pp)
MinLoss				
MLVU	Order	25.71	31.43	*+5.71*
MVBench	Character Order	67.50	74.50	*+7.00*
VideoMME	Temporal Perception	67.27	76.36	*+9.09*
MVBench	Moving Count	53.00	63.00	*+10.00*
MLVU	SportsQA	36.11	47.22	***+11.11***
Diverse				
MLVU	SportsQA	36.11	44.44	*+8.33*
MVBench	State Change	61.00	70.50	*+9.50*
MLVU	Order	25.71	35.71	*+10.00*
VideoMME	Temporal Perception	67.27	78.18	*+10.91*
MVBench	Moving Count	53.00	66.00	***+13.00***
Temp				
MLVU	PlotQA	44.00	48.00	*+4.00*
MLVU	Order	25.71	30.00	*+4.29*
MLVU	SportsQA	36.11	41.67	*+5.56*
VideoMME	Temporal Perception	67.27	74.55	*+7.27*
MVBench	Moving Count	53.00	63.00	***+10.00***
Temp+				
MVBench	Character Order	67.50	74.50	*+7.00*
MLVU	Order	25.71	32.86	*+7.14*
MVBench	Counterfactual Inference	59.50	68.00	*+8.50*
MLVU	SportsQA	36.11	47.22	*+11.11*
VideoMME	Temporal Perception	67.27	85.45	***+18.18***

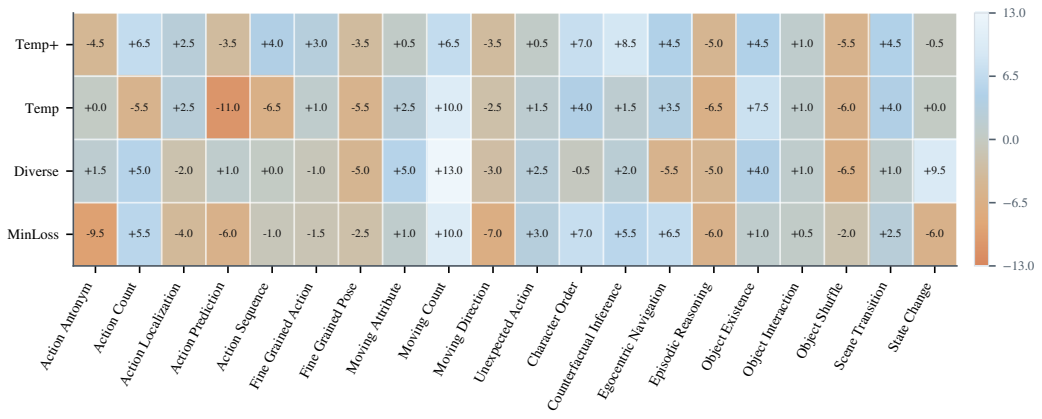


Figure 7: **MVBench Subtask Delta Heatmap.** Each cell shows the final displayed difference between a GDO profile and the frozen benchmark-specific UNI-10X reference on the same MVBench subtask. The heatmap displays gains, unchanged cells, and adverse outcomes across the full subtask map.

B SUBTASK ANALYSIS

Table 4 reports a post hoc descriptive ranking of the five temporally related subtasks with the largest gain for each profile. Recurring categories include temporal perception in VIDEOMME, moving-count and order-sensitive subtasks in MVBENCH, and Order and SportsQA in MLVU. The main-text table and heatmap retain adverse outcomes.

Figure 8 complements the final subtask deltas in the main paper by showing how representative subtasks evolve across checkpoints.

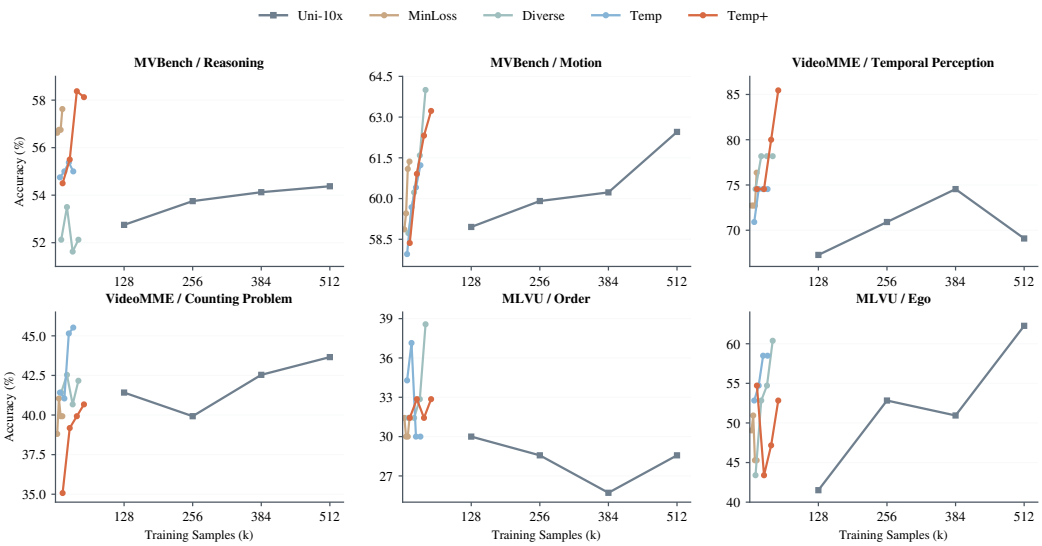


Figure 8: **Subtask Trajectories.** Panels show persistent temporal gains and adverse subtasks against the frozen benchmark-specific UNI-10X reference.

The trajectory view provides additional context for the final subtask table. VideoMME Temporal Perception and MLVU Order separate progressively as temporal pressure increases, while MVBench Reasoning shows that the gain is not restricted to one notion of motion. These trajectories show that several benchmark-level differences persist across recorded checkpoints and concentrate in identifiable temporal and reasoning subtasks. The adverse panels are equally informative. VideoMME Counting Problem and MLVU Ego do not follow the same progression, and at some checkpoints the UNI-10X control remains stronger. These cases show that the aggregate gains do not correspond to uniform improvements across all subtasks.

C METHOD DETAILS

C.1 SHARED SCORE

The released builder separates preference from admissibility. A shared score ranks candidates, and the goal profile determines which ranked candidates remain feasible under a given allocation target.

After descriptor extraction and merge, each sample keeps the descriptor vector

$$\left[m_{\text{flow}},\ m_{\text{vds}},\ m_{\text{tnc}},\ m_{\text{sc}},\ m_{\text{ppl}},\ m_{\text{cov}}\right],$$

and the merge stage also produces one scalar `quality_score`. The scorer then applies one ranking function to short-video samples and another to image-QA samples.

For short-video samples, the scorer uses

$$\rho_{\text{vid}}(x) = 0.35\,\tanh\left(\frac{b_{\text{vid}}(x)}{3}\right) + 0.95\,z_{\text{vds3}}(x) + 0.35\,z_{\text{qual}}(x), \tag{12}$$

$$b_{\text{vid}}(x) = q_{\text{text}}(x) + 0.85\,d(x) + 0.90\,a(x) + 0.55\,t(x) + 0.15\,r_{\text{src}}(x). \tag{13}$$

q_{text} is a heuristic question/answer quality prior, d is a medium-difficulty preference, a is a bin-alignment term over duration bucket, temporal bucket, question form, question length, answer length, and source type, t is the short-video temporal bonus, r_{src} is a source-rarity prior, z_{vds3} is the normalized video-dependence term, and z_{qual} is the normalized merged `quality_score`. For

Table 5: **Temp+ Ablation.** The first row is the full Temp+ builder. Each subsequent row removes one or more score contributions; parentheses report the pp change relative to full Temp+.

Ablation	MVBench	VideoMME	MLVU
Temp+	63.65	62.89	46.89
Temp+ w/o VDS	63.12 *(-0.53)*	61.78 *(-1.11)*	47.60 *(+0.71)*
Temp+ w/o PPL	63.30 *(-0.35)*	61.81 *(-1.07)*	47.00 *(+0.11)*
Temp+ w/o SC	62.25 *(-1.40)*	62.15 *(-0.74)*	43.73 *(-3.16)*
Temp+ w/o VDS/PPL/SC	61.65 *(-2.00)*	62.04 *(-0.85)*	47.06 *(+0.17)*

image QA, the builder uses

$$\rho_{\text{img}}(x) = 0.90 \tanh\left(\frac{b_{\text{img}}(x)}{3}\right) + 0.15\, z_{\text{qual}}(x), \tag{14}$$

$$b_{\text{img}}(x) = 1.10\, q_{\text{text}}(x) + 0.85\, d(x) + 0.90\, a(x) + 0.15\, r_{\text{src}}(x). \tag{15}$$

The image score omits video-only terms such as the temporal bonus and the VDS3 addition. Thus, the short-video score combines a heuristic base term, a normalized video-dependence term, and the merged quality term, while the image score follows the same logic without the video-only additions. Because the terms are normalized before mixing, the coefficients are fixed relative weights rather than incomparable raw scales. All profiles use the same scorer; only the goal-specific feasibility controls change. The profile comparison therefore isolates changes in data allocation from changes to the scoring function.

C.2 Goal Profiles

Table 6 makes the four released goal profiles explicit. The shared score is fixed, while the budget and feasibility controls change across profiles.

Table 6: **Goal Profiles.** N_g is the final $1\times$ subset size, r_v is the selected video ratio, and the remaining columns are the main feasibility controls.

Goal	N_g	r_v	VDS+ tgt	$r_v^{\min}$	$r_v^{\max}$	$r_{t\|v}^{\min}$
MinLoss	12.9k	0.32	2600	0.15	0.32	0.05
Diverse	42.9k	0.45	5000	0.25	0.45	0.15
Temp	33.3k	0.50	6500	0.35	0.50	0.20
Temp+	53.3k	0.59	9000	0.50	0.64	0.38

N_g denotes the final $1\times$ subset size used for SFT, r_v the selected video ratio, VDS+ tgt the target count of VDS-positive samples used during budget construction, and $r_{t|v}^{\min}$ the minimum temporal-positive ratio within selected videos. MinLoss targets low-loss supervision under the smallest budget. Diverse restores broader semantic and source coverage. Temp and Temp+ raise both the selected video ratio and the temporal-positive floor. The benchmark and subtask analyses compare the outcomes associated with these preset changes under the fixed training protocol.

D Limitations

LVBench provides the main benchmark-level limitation. It targets ultra-long videos that are weakly represented in the short-video LLaVA-Video pool, while the training mixture also contains substantial LLaVA-OneVision image QA. Its gain is therefore smaller than the gains on MVBench and MLVU, although the scores still improve from MinLoss to Temp and Temp+. The temporal-necessity signal is also only a question-level proxy, and the released profiles vary feasibility constraints more than the shared score itself. Stronger goal-specific scorers and native ultra-long-video pools remain open directions.

E QUALITATIVE EXAMPLES

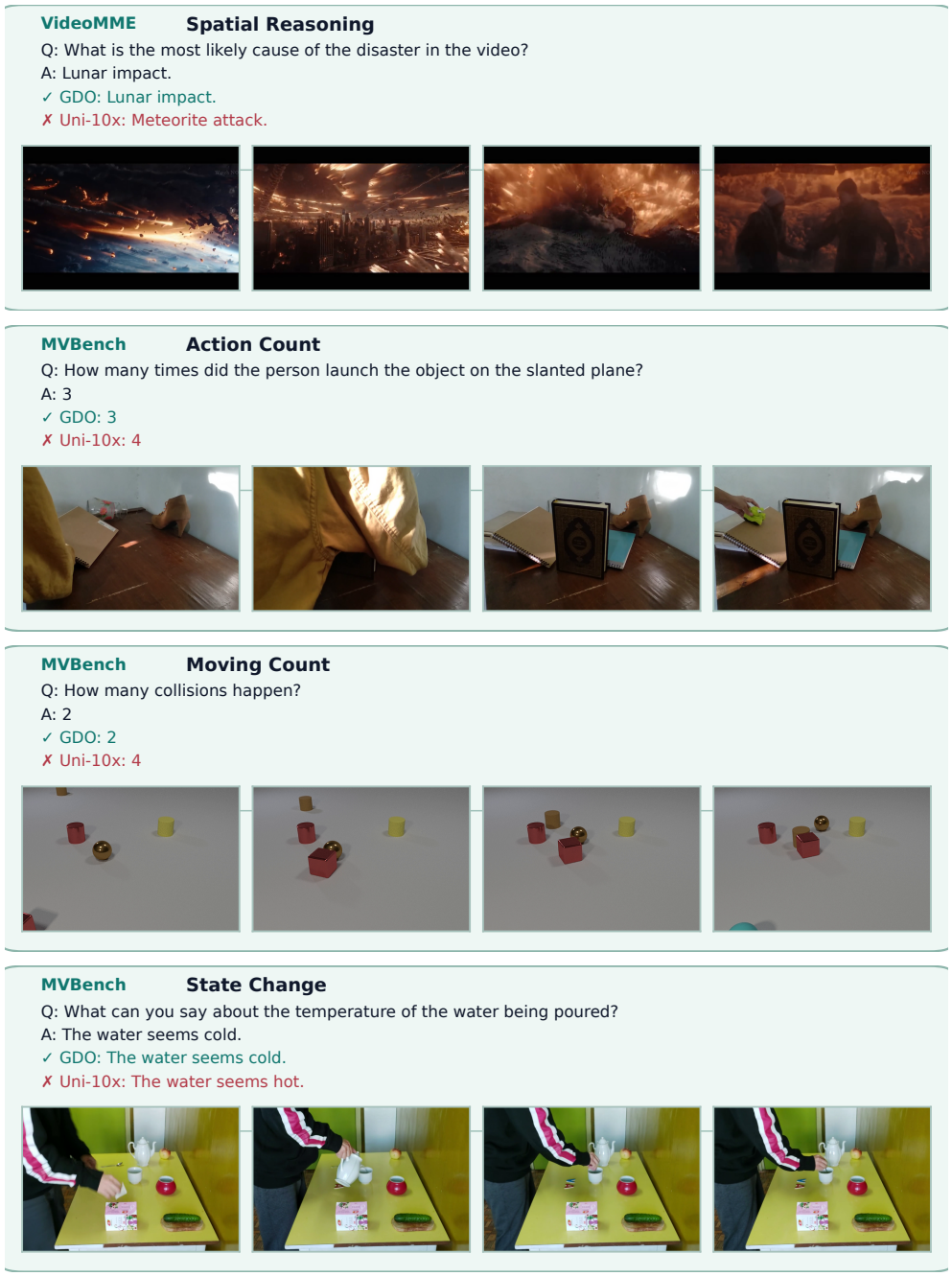
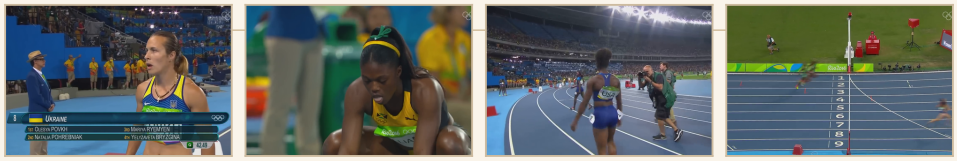


Figure 9: **GDO Wins.** Representative cases where GDO resolves the decisive cue before UNI-10X, including causal interpretation in VIDEOMME, repeated-action counting in MVBENCH Action Count, collision counting in Moving Count, and temperature-based state inference in State Change. Each row shows sampled frames, the question, the answer, and both model outputs.

VideoMME **Object Recognition**

Q: Which athlete is the second to reach the finish line in the video?
A: The athlete wearing a yellow top and yellow shorts.
✗ GDO: The athlete wearing a white top and yellow shorts.
✗ Uni-10x: The athlete wearing a white top and yellow shorts.



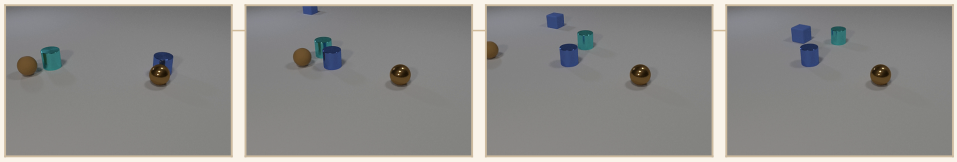
MVBench **State Change**

Q: What can you say about the temperature of the water being poured?
A: The water seems cold.
✗ GDO: The water seems hot.
✗ Uni-10x: The water seems hot.



MVBench **Moving Count**

Q: How many cylinders enter the scene?
A: 0
✗ GDO: 1
✗ Uni-10x: 2



MVBench **Action Prediction**

Q: What will the person do next?
A: Take the blanket.
✗ GDO: Put down the food.
✗ Uni-10x: Put down the food.

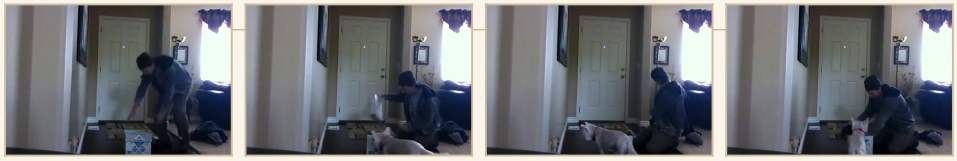


Figure 10: **Difficult Cases for Both.** Cases that remain hard for both models under the fixed train/eval contract, including finish-order recognition in VIDEOMME, subtle temperature inference in State Change, small-object counting in Moving Count, and next-action prediction in Action Prediction. These rows make the remaining failure modes concrete.

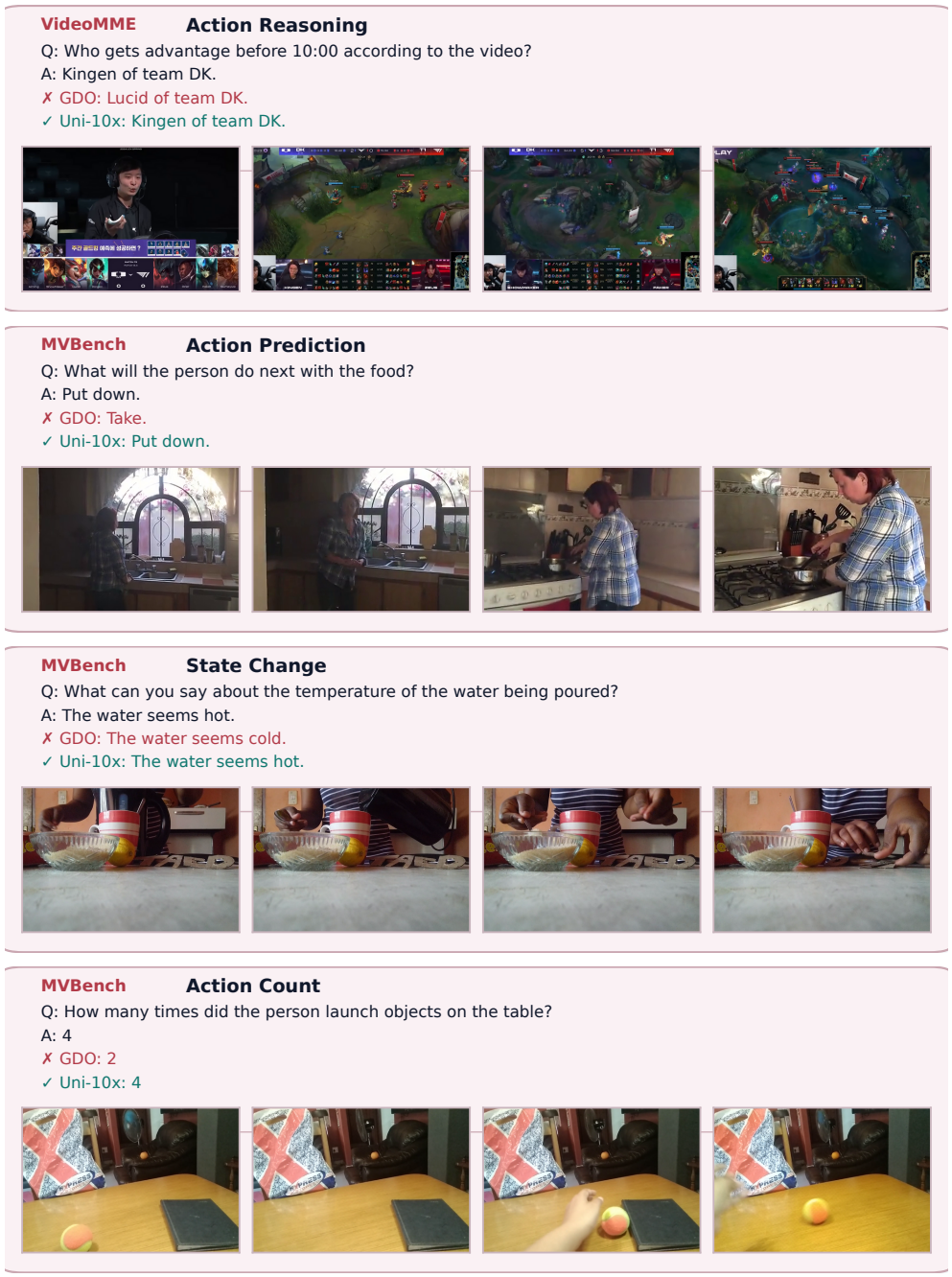


Figure 11: **UNI-10X Wins.** Cases where the frozen benchmark-specific UNI-10X reference remains better, including esports action reasoning in VIDEOMME, object-heavy next-action prediction, subtle water-temperature inference, and repeated-action counting. These rows show that the gain from temporal allocation is not uniform across all question types.