LOST IN ADAPTATION: LAYER-SELECTIVE RECOVERY OF TEMPORAL REASONING IN VIDEO-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal adaptation can weaken temporal reasoning (TR) in video-language models (VLMs) even when the visual evidence needed for an answer is correctly perceived. This weakened competence, however, remains partly recoverable from the model's paired pre-adaptation language checkpoint. We introduce MERIT, a gradient-free framework that selectively restores this competence across the language backbone while protecting temporal perception (TP). It assigns each self-attention layer a VLM-dominant or LLM-dominant interpolation and uses the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) to search the recipe space under a perception-constrained objective. Across three VLM families and five video benchmarks, MERIT improves reasoning-aligned performance in all 15 model–benchmark evaluations while preserving TP across all three Video-MME settings. Recipes selected using 232 diagnostic examples transfer to four unseen benchmarks, with relative gains up to 27.8%, and outperform full-backbone, uniform-attention, and alpha-tuned random-layer merging. Interventional masking shows that MERIT-selected layers are disproportionately important for TR, while frame-level attribution reveals increased reliance on temporally distributed, causally relevant evidence. These results show that adaptation-induced reasoning degradation can be repaired post hoc through selective reuse of pre-adaptation competence, without learning new parameters or adding inference overhead.[1]

1 INTRODUCTION

Multimodal adaptation equips pretrained language models with visual capabilities, but it can also weaken reasoning abilities acquired during language pretraining (Li et al., 2024b). We observe this regression in temporal reasoning (TR). As shown in Figure 1, the language backbone of an adapted video-language model (VLM) performs worse than its original LLM checkpoint on text-only TR, indicating that part of the degradation arises within the language component itself. In video settings, the same loss can appear as a gap between perception and reasoning: the model recognizes relevant objects and actions, yet fails to infer event order, state transitions, or causal dependencies. Such failures are particularly consequential for video temporal reasoning, which requires tracking temporal and causal structure (Chu et al., 2024; Wang & Zhao, 2024) and integrating evidence across frames (Chen et al., 2024; Fang et al., 2024; Nie et al., 2024; Cheng et al., 2025). This suggests that some video reasoning failures reflect capabilities weakened during multimodal adaptation, even when the relevant visual evidence remains available.

This observation suggests that temporal reasoning need not be learned anew: the original LLM checkpoint already provides a source of the competence weakened during multimodal adaptation. At the same time, the adapted VLM provides the perceptual grounding required for video understanding. The resulting challenge is to recover the former while preserving the latter. Fine-tuning can improve reasoning (Li et al., 2024b), but requires additional supervision and parameter updates. Model merging offers a direct post-hoc mechanism for recombining the complementary capabilities of the paired pre-adaptation and post-adaptation checkpoints. Prior work has explored evolutionary recipe search (Akiba et al., 2025), multimodal component integration (Sung et al., 2023; Zhu et al., 2025a),

[1] Code and data are available at: https://anonymous.4open.science/r/MERIT-ICLR27.

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.

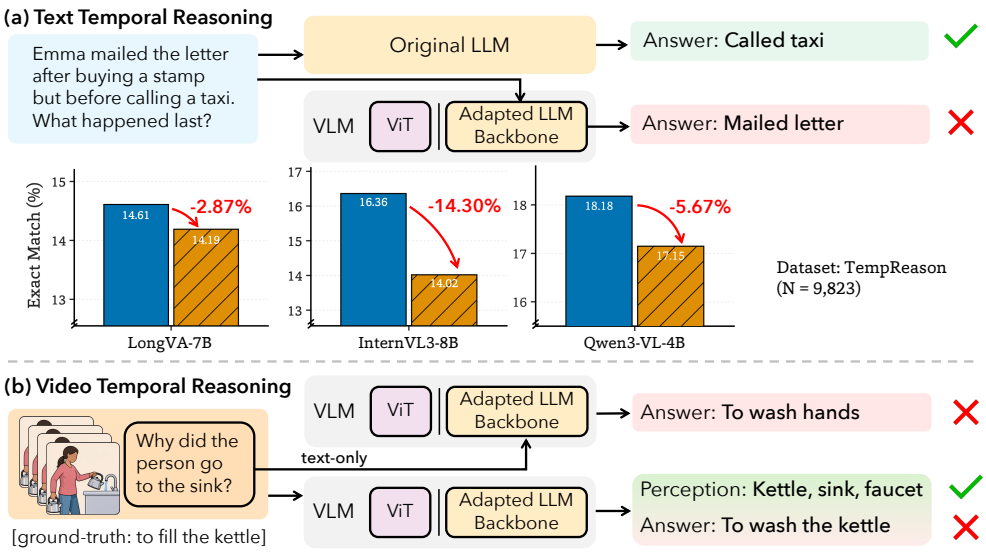

Figure 1: **Multimodal adaptation can preserve what a VLM sees while weakening how its language backbone reasons over time.** **(a)** The adapted language backbone underperforms its pre-adaptation LLM on text-only temporal reasoning (TR), directly isolating adaptation-induced degradation within the language component (§B). **(b)** In video TR, the VLM correctly perceives the relevant visual entities but fails to infer the temporal-causal structure required for the answer.

and reasoning transfer through merging (Chen et al., 2025). However, these approaches generally operate over entire backbones or predefined module groups, applying a similar compromise across the language stack. This granularity may be limiting, as reasoning behavior can be highly sensitive to particular layers or subnetworks (Sharma et al., 2024; Shi et al., 2024; Ju et al., 2024). The key challenge is therefore to determine where the original reasoning capability should be restored while leaving adaptation-induced perceptual competence intact. This leads to our central question:

Where should pre-adaptation reasoning competence be restored to recover temporal reasoning while preserving multimodal perception?

We introduce MERIT (*Merging for Enhanced Reasoning In Temporal Tasks*), a gradient-free framework for layer-selective capability recovery. Our key idea is to determine *where* the adapted VLM should retain multimodal parameters and where it should draw more strongly on the original LLM. Specifically, MERIT assigns each self-attention layer a VLM-dominant or LLM-dominant interpolation, producing a combinatorial space of merging recipes. Candidate recipes are evaluated with a perception-constrained objective that rewards TR improvement while penalizing temporal perception (TP) degradation. As this objective is non-differentiable with respect to discrete layer assignments, MERIT uses the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen & Ostermeier, 2001) to search the recipe space. All source parameters remain fixed throughout.

Across three VLM families and five video benchmarks, MERIT consistently recovers temporal reasoning while preserving perceptual capability. Recipes selected using only 232 diagnostic examples improve reasoning-aligned performance across all 15 model–benchmark comparisons and transfer to four unseen benchmarks, with relative gains of up to 27.8%. Evaluation on InternVL3-14B further shows that the same recovery pattern extends beyond the 4B–8B setting (§ D.4). Controlled comparisons show that the gains depend on the discovered layer configuration: MERIT outperforms globally tuned full-backbone merging, uniform attention merging, and alpha-tuned random layer selection. Although LoRA and full-parameter tuning fit the diagnostic examples more strongly, their gains transfer less reliably, consistent with MERIT recovering competence already encoded in the paired checkpoint. Interventional masking further shows that the selected layers are functionally important for temporal reasoning, while frame-level attribution indicates that the repaired models rely more strongly on temporally distributed and causally relevant evidence. These results establish adaptation-induced reasoning degradation as a tractable post-hoc repair problem and identify layer-selective merging as an effective mechanism for capability recovery.

2 Related Work

Video Temporal Reasoning. Video temporal reasoning requires modeling event order, state changes, and causal relations across frames. Benchmarks including CLEVRER (Yi et al., 2019), NExT-QA (Xiao et al., 2021), TempCompass (Liu et al., 2024), TOMATO (Shangguan et al., 2024), TVBench (Cores et al., 2024), Vinoground (Zhang et al., 2024a), and LongVideoBench (Wu et al., 2024) expose persistent weaknesses in temporal perception and reasoning. Li et al. (2024a) further show that the language backbone can bottleneck video TR even when visual representations contain sufficient temporal information. MERIT identifies a more specific source of this bottleneck: multimodal adaptation can weaken temporal competence already present in the pre-adaptation language backbone.

Improving Reasoning in Video-Language Models. Training-based methods improve temporal reasoning through time-aware architectures and instruction tuning (Ren et al., 2024; Huang et al., 2024), self-training and distillation (Shi et al., 2025; Qiu et al., 2025; Fei et al., 2024), or text-only temporal supervision (Li et al., 2024a). Training-free methods modify inference through task decomposition (Min et al., 2024; Liao et al., 2025), constrained decoding (Qi et al., 2025), or reasoning demonstrations (Wang et al., 2025a). MERIT instead performs post-hoc parameter-level recovery using paired checkpoints, without gradient updates or additional inference procedures.

Adaptive Model Merging. Model merging combines capabilities across checkpoints without conventional retraining. AdaMerging (Yang et al., 2024b) learns task- or layer-wise merging coefficients, while evolutionary model merging (Akiba et al., 2025) searches merging recipes through black-box optimization. MERIT applies evolutionary search to a different objective: selecting layer-wise interventions between paired pre- and post-adaptation checkpoints while protecting temporal perception.

Multimodal Model Merging and Capability Recovery. Prior multimodal merging work combines vision, language, and cross-modal components (Sung et al., 2023), allocates reusable multimodal modules (Zhu et al., 2025a), or injects reasoning from specialized LLMs into VLMs (Chen et al., 2025). Locate-then-Merge (Yu & Ananiadou, 2025) addresses multimodal catastrophic forgetting through neuron-level parameter fusion, while recent methods use adaptive cross-layer matching (Zhang et al., 2026) or weight-decomposed merging (Zou et al., 2026). MERIT specifically recovers adaptation-induced temporal reasoning loss by searching for task-effective self-attention interventions in the VLM's own pre-adaptation backbone under an explicit perception-preservation objective.

3 MERIT: Layer-Selective Temporal Reasoning Recovery in VLMs

MERIT treats adaptation-induced reasoning degradation as a selective recovery problem. We assume access to an adapted VLM and its paired pre-adaptation language backbone. The two checkpoints contain complementary capabilities: the adapted model provides the visual grounding needed for video understanding, while the earlier language checkpoint retains reasoning competence weakened during multimodal adaptation. Globally moving toward either checkpoint can therefore trade reasoning against perception. MERIT instead searches for a structured combination in which different self-attention layers draw to different degrees from the two checkpoints. The resulting recipe is selected to improve temporal reasoning (TR) while explicitly protecting temporal perception (TP), as illustrated in Figure 2.

3.1 Recovery Setting and Perception-Constrained Objective

Let M denote an adapted VLM and N its paired pre-adaptation text-only LLM, with a shared language architecture. This pairing isolates a single adaptation trajectory: N provides the parameter state before multimodal adaptation, while M contains the capabilities acquired afterward. Let g denote a merging recipe that determines how the self-attention parameters of the two checkpoints are combined across the L language layers. We parameterize g explicitly in § 3.2. Let $\mathrm{Acc}_t(g)$ denote the accuracy of the resulting model on task $t \in \{\mathrm{TP}, \mathrm{TR}\}$.

The desired repair is asymmetric. We seek to recover TR weakened during adaptation while avoiding loss of the TP competence already present in M. Therefore, we define perceptual degradation relative to the adapted VLM:

$$D_{\mathrm{TP}}(g) = \max\left(0,\ \mathrm{Acc}_{\mathrm{TP}}^{\mathrm{base}} - \mathrm{Acc}_{\mathrm{TP}}(g)\right), \tag{1}$$

where $\mathrm{Acc}_{\mathrm{TP}}^{\mathrm{base}}$ is the TP accuracy of M. The one-sided form penalizes only degradation relative to the adapted model; a recipe that also improves perception is not discouraged.

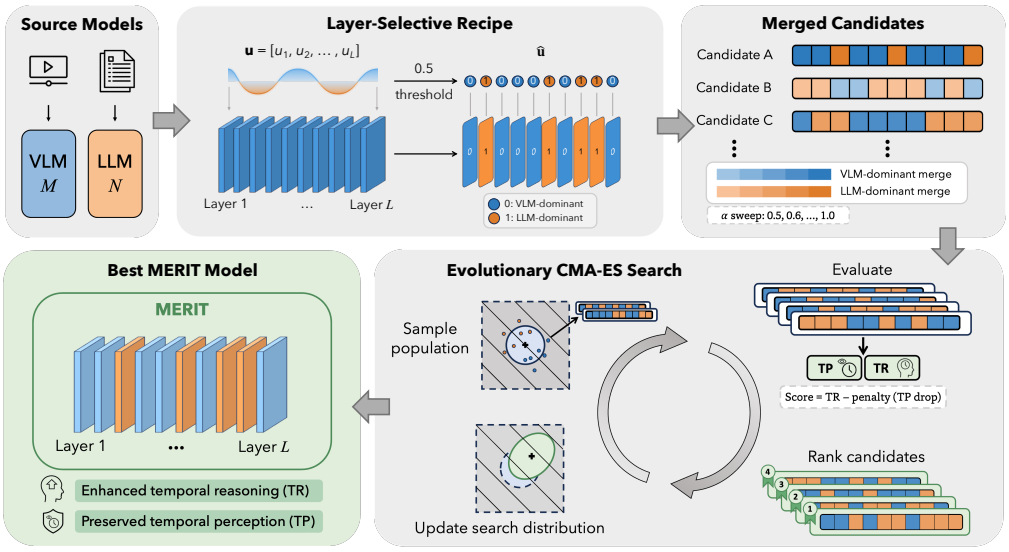

Figure 2: **Overview of MERIT.** Given an adapted VLM and its paired pre-adaptation LLM, MERIT searches for a layer-wise self-attention merging recipe in which each layer is VLM-dominant or LLM-dominant. CMA-ES optimizes this discrete structure under a perception-constrained objective that rewards TR recovery while penalizing TP degradation. The selected recipe is compiled into a single merged VLM with no additional inference components.

We then optimize

$$g^* = \arg\max_g \mathcal{F}(g), \qquad \mathcal{F}(g) = \mathrm{Acc}_{\mathrm{TR}}(g) - \lambda D_{\mathrm{TP}}(g), \tag{2}$$

where λ controls the strength of perceptual protection. This objective directly encodes the recovery criterion: a useful intervention should improve the weakened capability while preserving what multimodal adaptation added.

Importantly, TP and TR examples provide only black-box signals for selecting among candidate recipes. They do not update either checkpoint. Improvements must therefore arise by recombining competence already present in M and N, instead of by fitting new parameters to the diagnostic set.

3.2 Layer-Selective Self-Attention Merging

A central question is where the pre-adaptation checkpoint should be reintroduced. A single global interpolation forces every layer to make the same trade-off between the two checkpoints, even though multimodal adaptation need not affect all regions of the language backbone equally. We therefore expose the layer structure itself as the decision space: some layers can remain close to the adapted VLM, while others can draw more strongly from the pre-adaptation LLM.

We intervene on self-attention because temporal reasoning requires selecting question-relevant evidence and integrating information across temporally separated visual tokens. Self-attention directly mediates these interactions, making it a natural target for selective recovery. Within self-attention, we use layers as the unit of selection to avoid the coarse granularity of global merging without incurring the large, architecture-specific search space of head-level selection. Layer-level recipes remain portable across model families while still allowing different regions of the backbone to draw on different capability sources. Notably, this formulation does not assume that TR is localized to individual layers; it asks which layer-level interventions provide effective recovery with limited perceptual interference.

We instantiate each recipe as $g = (\hat{\mathbf{u}}, \alpha)$, where $\hat{\mathbf{u}} \in \{0,1\}^L$ specifies the layer-wise dominance pattern and α controls interpolation strength. To search over the discrete pattern, we parameterize it through a continuous gating vector $\mathbf{u} \in [0,1]^L$ and threshold each coordinate:

$$\hat{u}_\ell = \begin{cases} 1, & u_\ell \geq 0.5, \\ 0, & u_\ell < 0.5. \end{cases} \tag{3}$$

Given $\hat{\mathbf{u}}$, the self-attention parameters at layer ℓ are constructed as

$$\theta_\ell = \begin{cases} (1-\alpha)\theta_\ell^M + \alpha\theta_\ell^N, & \hat{u}_\ell = 1, \\ \alpha\theta_\ell^M + (1-\alpha)\theta_\ell^N, & \hat{u}_\ell = 0. \end{cases} \tag{4}$$

The interpolation is defined for $\alpha \in [0, 1]$. Equation 4 has an exact complement symmetry: $(\hat{\mathbf{u}}, \alpha)$ and $(\mathbf{1} - \hat{\mathbf{u}}, 1 - \alpha)$ produce the same merged checkpoint. Thus, when layer assignments are allowed to vary, coefficients below 0.5 introduce no additional checkpoints, and we restrict the search to $\alpha \in [0.5, 1]$ without loss of expressivity. Within this range, $\hat{u}_\ell = 1$ denotes an LLM-dominant layer and $\hat{u}_\ell = 0$ a VLM-dominant layer. At $\alpha = 0.5$, the two assignments coincide.

This parameterization separates two decisions that global merging conflates: *where* the pre-adaptation checkpoint should dominate and *how strongly* the two checkpoints should be mixed. It therefore allows MERIT to preserve multimodal adaptation where it remains useful while restoring pre-adaptation parameters where they improve reasoning. We evaluate self-attention against MLP-only, attention-plus-MLP, and full-block interventions in § 4.3.

3.3 Evolutionary Recipe Search

Layer-selective recovery creates a combinatorial black-box optimization problem. For fixed $\alpha \neq 0.5$, L self-attention layers induce 2^L binary layer patterns; a 28-layer backbone already yields more than 268 million candidates. Moreover, layer decisions can interact, so the value of one assignment cannot be assessed independently of the others. Recipe quality is observable only after constructing the merged checkpoint and evaluating downstream TR and TP.

We optimize this objective using the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen & Ostermeier, 2001). For a fixed α, CMA-ES maintains a distribution over continuous gating vectors $\mathbf{u}$, thresholds them into binary patterns $\hat{\mathbf{u}}$, instantiates the corresponding recipes, and evaluates their fitness with $\mathcal{F}$. Its adaptive covariance captures dependencies among layer decisions, enabling structured exploration of interacting interventions.

Exploiting the complement symmetry of Eq. 4, we search only one half of the interpolation range. We run an independent CMA-ES search for each $\alpha \in \{1.0, 0.9, 0.8, 0.7, 0.6\}$. At $\alpha = 0.5$, the two branches of Eq. 4 coincide, so the layer pattern has no effect and only one recipe is evaluated. We retain the recipe with the highest objective across all candidates. The outer grid controls mixing strength, while CMA-ES determines where each checkpoint should dominate. This separation also allows us to test whether gains arise from the discovered layer structure or merely from interpolation strength.

Once selected, the recipe is compiled directly into a single VLM checkpoint. MERIT therefore incurs an offline search cost while leaving deployment unchanged: the recovered model has the same architecture and inference procedure as the original VLM, with no trainable adapters, ensembles, or additional decoding steps.

4 Experiments

We evaluate MERIT through four questions about adaptation-induced capability recovery:

- **RQ1: Recovery and transfer (§ 4.2).** Is adaptation-induced TR loss recoverable while preserving TP, and does this recovery generalize beyond the diagnostic set?
- **RQ2: Determinants of recovery (§ 4.3).** Does successful recovery depend on *where* pre-adaptation parameters are restored, beyond interpolation strength or supervised fitting?
- **RQ3: Functional intervention (§ 4.4).** Are the selected layers especially effective intervention points for temporal reasoning?
- **RQ4: Temporal grounding (§ 4.5).** Does recovery shift decisions toward temporally distributed, causally informative evidence?

4.1 Experimental Setup

We evaluate MERIT on LongVA-7B (Zhang et al., 2024b), InternVL3-8B (Zhu et al., 2025b), and Qwen3-VL-4B-Instruct (Bai et al., 2025). Each VLM is paired with the pre-adaptation language backbone specified by its original release. Exact checkpoints and architectures are provided in § C.1.

Search and Evaluation Data. We use 55 Temporal Perception (TP) and 177 Temporal Reasoning (TR) examples from Video-MME (Fu et al., 2025) solely for recipe selection and in-domain diagnostic evaluation. The remaining 2,468 Video-MME examples form the held-out Others subset. Once selected, each recipe is frozen and evaluated on four unseen benchmarks: LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2025b), MMBench-Video (Fang et al., 2024), and Video-Holmes (Cheng et al., 2025). This separation tests whether a compact diagnostic signal can identify transferable recovery interventions. Dataset and metric details are provided in § C.2 and § F.1.

Compared Methods. We compare MERIT with the unmodified VLM (Base); supervised LoRA and full-parameter fine-tuning (FFT) on the same 232 examples; Full-Backbone and All-Attention interpolation; and Random-k with either fixed or diagnostic-tuned interpolation strength. For both random-layer baselines, k matches the number of layers selected by MERIT, and we report mean and standard deviation over five runs. Full configurations are provided in § C.3.

Search and Evaluation Protocol. We optimize layer-wise recipes with CMA-ES (Hansen & Ostermeier, 2001) using `mergekit-evolve` (Goddard et al., 2024). Each fixed-α search evaluates at most 1,600 candidates and retains the recipe with the highest perception-constrained objective. All source parameters remain fixed, and evaluation uses deterministic decoding and official benchmark protocols. Further implementation details are provided in § C.4.

4.2 TR Recovery Generalizes Beyond the Diagnostic Set

Tables 1 and 2 test whether adaptation-induced TR degradation can be recovered without sacrificing the perceptual capability acquired during multimodal adaptation.

Table 1: **MERIT recovers temporal reasoning while preserving perception on Video-MME.** Accuracy (%) on the TP/TR diagnostic subsets and held-out Others. Shaded entries are in-sample; † denotes α-tuned Random-k; parentheses show relative change over Base.

Model	Variant	Diagnostic		Held-out
		TP	TR	Others
LongVA-7B	Base	60.0	40.1	55.0
	LoRA‡	72.7	53.1	54.7
	FFT‡	81.8	64.4	53.0
	Full-Backbone	60.0	40.1	55.0
	All-Attention	**61.8**	39.5	55.2
	Random-k	58.9±1.6	38.4±1.7	54.7±0.5
	Random-k†	59.6±0.8	40.8±0.9	54.8±0.5
	MERIT	**61.8** (+3.0%)	**49.7** (+23.9%)	**55.3** (+0.5%)
InternVL3-8B	Base	74.5	52.0	67.4
	LoRA‡	83.6	71.2	67.4
	FFT‡	85.5	71.2	66.3
	Full-Backbone	76.4	53.1	67.3
	All-Attention	76.4	54.2	67.3
	Random-k	78.2±2.9	51.6±0.9	67.5±0.3
	Random-k†	78.6±2.3	53.1±0.9	67.3±0.3
	MERIT	**81.8** (+9.8%)	**57.6** (+10.8%)	**67.9** (+0.7%)
Qwen3-VL-4B	Base	69.1	45.2	65.3
	LoRA‡	76.4	57.1	65.0
	FFT‡	85.5	70.6	64.1
	Full-Backbone	67.3	45.8	64.7
	All-Attention	67.3	44.1	**65.5**
	Random-k	63.3±3.0	41.5±2.7	62.7±1.2
	Random-k†	69.5±1.5	44.4±1.1	62.4±2.6
	MERIT	**70.9** (+2.6%)	**46.9** (+3.8%)	64.6 (-1.1%)

Recovery transfers without undoing multimodal adaptation. On Video-MME, MERIT improves TR by 23.9%, 10.8%, and 3.8% for LongVA-7B, InternVL3-8B, and Qwen3-VL-4B, respectively, while preserving or improving TP in all three models. The held-out Others subset changes by only $+0.5\%$, $+0.7\%$, and -1.1%, indicating limited impact on non-target capabilities. More importantly, after selection on only 232 Video-MME examples, each recipe remains fixed and improves reasoning-aligned performance over Base in all 12 comparisons across four unseen benchmarks. Including Video-MME TR, MERIT improves all 15 reasoning-aligned evaluations, with relative gains reaching 27.8% on MMBench-Video TR and 24.3% on Video-Holmes. Thus, the selected interventions recover reasoning behavior that transfers across datasets without broadly erasing the capabilities acquired during adaptation.

The same recovery appears at the language-backbone level. On TempReason, multimodal adaptation reduces text-only TR across all three model families, and MERIT consistently restores part of this loss (§ B). This closes the loop between diagnosis and recovery: the same intervention

Table 2: **MERIT transfers temporal reasoning gains beyond the diagnostic set.** Fixed recipes selected on Video-MME are evaluated on four unseen video benchmarks. Parentheses show relative change over BASE; † denotes α-tuned RANDOM-K. MMBench-Video uses an LLM-as-a-judge score; all other results are accuracy (%). Best performance is marked in bold.

Model	Variant	LongVideoBench		LVBench		MMBench-Video		Video-Holmes
		Relation	Overall	TR	Overall	TR	Overall	Overall
LongVA-7B	BASE	48.0	52.5	45.3	37.6	1.05	1.27	15.2
	LORA	47.2	52.2	45.3	37.7	1.04	1.24	14.5
	FFT	35.5	40.8	41.8	35.4	0.98	1.15	4.7
	FULL-BACKBONE	47.9	52.9	44.8	38.1	1.08	1.25	14.6
	ALL-ATTENTION	48.0	52.6	46.3	38.0	1.08	1.28	15.4
	RANDOM-K	48.1±0.7	52.7±0.2	44.4±1.6	37.8±0.6	1.07±0.02	1.25±0.02	14.5±3.8
	RANDOM-K†	48.2±0.7	52.9±0.3	45.3±1.8	37.7±0.8	1.07±0.01	1.25±0.02	15.8±2.5
	MERIT	**48.6** (+1.3%)	**53.6** (+2.1%)	**46.8** (+3.3%)	**39.7** (+5.6%)	**1.09** (+3.8%)	**1.29** (+1.6%)	**18.9** (+24.3%)
InternVL3-8B	BASE	54.8	59.8	51.2	44.8	1.44	1.68	36.5
	LORA	54.4	58.9	52.2	44.7	1.42	1.71	36.2
	FFT	52.2	55.7	50.2	44.4	1.41	1.69	35.7
	FULL-BACKBONE	55.1	59.9	50.7	44.7	1.44	1.71	36.3
	ALL-ATTENTION	54.8	59.5	50.2	44.7	1.44	1.71	36.2
	RANDOM-K	55.1±0.8	59.5±0.5	49.8±1.3	44.2±0.7	1.43±0.02	1.70±0.01	36.7±1.0
	RANDOM-K†	**55.2±0.5**	60.0±0.2	48.1±0.8	44.3±0.3	1.44±0.02	1.70±0.01	36.5±0.5
	MERIT	**55.2** (+0.7%)	**60.3** (+0.8%)	**53.2** (+3.9%)	**45.0** (+0.4%)	**1.46** (+1.4%)	**1.72** (+2.4%)	**37.2** (+1.9%)
Qwen3-VL-4B	BASE	56.0	61.9	39.3	39.4	1.08	1.50	31.5
	LORA	55.5	61.1	42.3	40.4	1.24	1.56	32.0
	FFT	51.5	55.3	43.3	40.3	1.11	1.53	30.7
	FULL-BACKBONE	55.9	60.4	33.3	34.9	1.26	1.56	31.2
	ALL-ATTENTION	56.3	**62.0**	43.3	**43.7**	1.27	1.59	31.9
	RANDOM-K	55.2±0.8	59.8±0.4	40.5±2.9	39.0±1.5	1.27±0.11	1.55±0.08	31.0±0.8
	RANDOM-K†	56.2±0.3	60.6±0.4	41.5±1.4	40.0±1.0	1.26±0.08	1.55±0.11	31.7±0.7
	MERIT	**56.9** (+1.6%)	60.6 (-2.1%)	**46.3** (+17.8%)	40.6 (+3.0%)	**1.38** (+27.8%)	**1.66** (+10.7%)	**32.9** (+4.4%)

that improves video reasoning also repairs the text-only temporal competence weakened during adaptation. An additional evaluation on InternVL3-14B shows the same recovery pattern beyond the 4B–8B setting (§ D.4). Selected recipes are reported in § D.1.

4.3 SUCCESSFUL RECOVERY DEPENDS ON WHERE AND HOW PARAMETERS ARE RESTORED

We next ask what makes recovery transferable, isolating the roles of intervention structure, search strategy, and supervised fitting.

Recovery depends on both where and what is restored. As shown in Tables 1 and 2, we tune α independently for FULL-BACKBONE, ALL-ATTENTION, and RANDOM-K on the same TP/TR diagnostic set used by MERIT. These baselines remain below MERIT on reasoning-aligned evaluations even after controlling for interpolation strength, showing that the gains cannot be explained by a favorable mixing coefficient alone. We further fix the MERIT layer pattern and compare attention-only merging with MLP-only, attention-plus-MLP, and full-block interventions. Table 6 shows that attention-only merging provides the strongest and most transferable recovery across Video-MME, LVBench, and Video-Holmes. Effective recovery therefore depends on both *where* parameters are restored and *which* parameters are modified within those layers.

Structured search finds better recovery recipes. Figure 6 compares CMA-ES with uniform random search under the same α, recipe space, objective, and 1,600-candidate budget. CMA-ES reaches higher final objective values on all three VLMs, and its independent runs converge to the same final objective and recipe. On InternVL3-8B, Table 4 further shows that the best random-search recipes improve Video-MME TR but remain below MERIT on both the diagnostic task and unseen benchmarks. Structured exploration of dependencies among layer assignments therefore matters even under a matched search space and evaluation budget.

Better fitting does not imply better recovery. Tables 1 and 2 show that LORA and FFT achieve substantially higher TP/TR accuracy on the 232 optimization examples, yet their gains transfer less reliably to held-out categories and unseen benchmarks. MERIT can only recombine parameters already present in the paired checkpoints. This gap between in-sample fit and out-of-sample transfer is consistent with MERIT selecting reusable competence already encoded in the model, while supervised tuning can specialize more strongly to the diagnostic set.

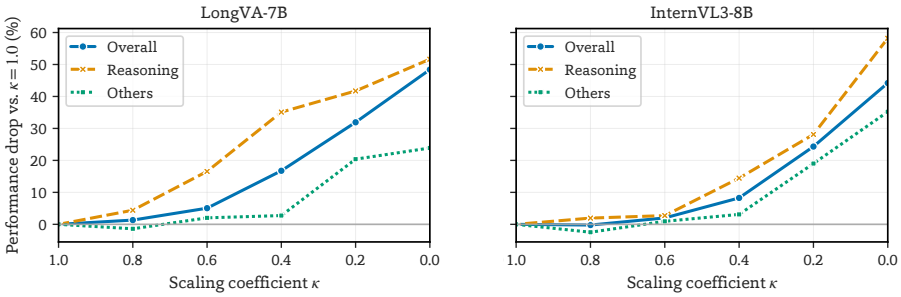

Figure 3: **MERIT selects reasoning-sensitive intervention points.** Suppressing the self-attention outputs of the selected layers in the unmodified VLM causes *Reasoning* to degrade markedly faster than *Overall* and *Others*.

A fixed-α search on InternVL3-8B evaluates at most 1,600 candidates and takes approximately 1.7 hours using 4 GPUs and 16 parallel workers (Table 7). As described in § 3.3, the complete grid contains at most 8,001 candidate evaluations ($5 \times 1{,}600$ search evaluations plus one evaluation at $\alpha = 0.5$), and the five searches can be parallelized across α. Once selected, the recipe is compiled into the checkpoint and adds no trainable parameters or inference overhead. Full search comparisons and cost details are provided in § D.1 and § D.3.

4.4 Selected Layers Are Reasoning-Sensitive Intervention Points

The previous results show that layer identity matters for recovery. We next ask whether the selected locations also have a distinct functional relationship with temporal reasoning. To separate this question from merging itself, we intervene directly on the unmodified VLM by suppressing the self-attention outputs of the selected layers.

For each selected layer, we scale its self-attention output by $\kappa \in [0, 1]$, where $\kappa = 1$ leaves the output unchanged and $\kappa = 0$ fully suppresses it while preserving the residual connection. We evaluate LVBench at $\kappa \in \{1.0, 0.8, 0.6, 0.4, 0.2, 0.0\}$ and define

$$\Delta(\kappa) = \frac{\text{Score}(1.0) - \text{Score}(\kappa)}{\text{Score}(1.0)}. \tag{5}$$

We compare **Reasoning**, **Overall**, and **Others**, where **Others** is the mean of all non-reasoning sub-metrics.

Figure 3 shows a clear separation. For LongVA-7B at $\kappa = 0.4$, **Reasoning** drops by 35.1%, compared with 16.8% for **Overall** and 2.7% for **Others**; complete suppression raises the reasoning drop to 51.7%. InternVL3-8B exhibits the same pattern: at $\kappa = 0$, the corresponding drops are 58.2%, 44.2%, and 35.2%. The layers discovered through recovery search therefore coincide with intervention points to which temporal reasoning is especially sensitive.

4.5 Qualitative Analysis: Recovery Changes Temporal Grounding

Accuracy alone does not reveal whether recovery changes the evidence underlying a model's decision. We therefore examine how MERIT redistributes visual support across time. Using temporal-reasoning questions from Video-Holmes and InternVL3-8B as a representative model, we combine frame-level attribution at the answer step with generated reasoning traces.

Attributing the final decision. For an input x and generated output y formatted as `<think>...</think>\n<answer>...</answer>`, we compute attribution at the prediction step for the answer-option token. Let t denote this token's output position, $y_{<t}$ the generated prefix ending immediately before it, and z_t the next-token logits obtained under teacher forcing on $(x, y_{<t})$. Let $\mathcal{C}$ denote the set of candidate answer-option tokens and $c \in \mathcal{C}$ the model-selected option. Restricting z_t to $\mathcal{C}$, we define the attribution target as the logit margin between c and its strongest alternative:

$$s = z_t^{\mathcal{C}}(c) - \max_{c' \in \mathcal{C} \setminus \{c\}} z_t^{\mathcal{C}}(c'). \tag{6}$$

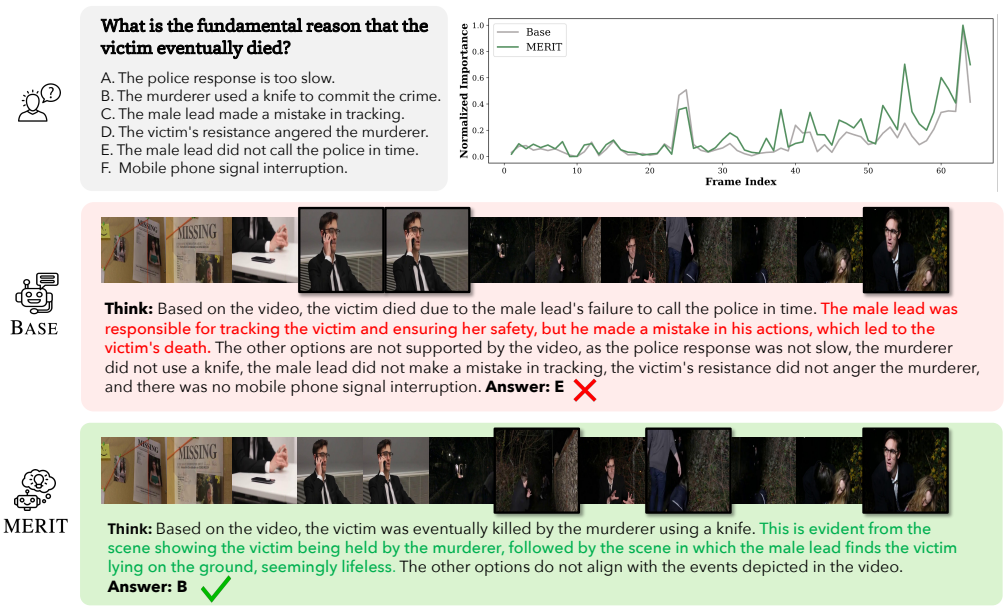

Figure 4: **MERIT shifts temporal grounding from a salient local cue to distributed causal evidence.** In this Video-Holmes example (Question ID: 463), BASE anchors on the phone-call scene and incorrectly attributes the victim's death to delayed help. MERIT instead assigns greater support to two temporally separated events, the victim being restrained and later found motionless, and recovers the correct causal explanation.

Frame-level evidence. Suppose the input contains T sampled frames. Let $V_\tau \in \mathbb{R}^{K\times d}$ denote the K projected visual-token embeddings of frame τ. We aggregate absolute gradient–activation products to obtain a non-negative attribution score for each frame:

$$I_\tau = \sum_{k=1}^{K}\sum_{j=1}^{d}\left|\frac{\partial s}{\partial V_\tau[k,j]}\cdot V_\tau[k,j]\right|. \tag{7}$$

For each model and input, we min–max normalize $\{I_\tau\}_{\tau=1}^{T}$ across frames to obtain a relative temporal attribution profile.

Recovery can shift support beyond salient local cues. Figure 4 shows a representative Video-Holmes example. The base model concentrates on a salient phone-call scene and incorrectly attributes the victim's death to delayed help. MERIT instead assigns greater importance to two temporally separated events: the victim being restrained and later found motionless. Linking these observations supports the correct causal explanation even though the killing itself is not explicitly shown.

This case illustrates a qualitative change beyond answer accuracy: the repaired model shifts support away from a misleading local cue and toward temporally separated events that jointly determine the causal explanation. Additional examples in § D.5 show related changes, including cases where both models reach the same answer but MERIT grounds it in a longer temporal event structure.

5 CONCLUSION

We study temporal reasoning degradation in VLMs as a capability recovery problem. MERIT selectively reuses the paired pre-adaptation language checkpoint to restore reasoning while preserving multimodal perception. Across three VLM families, compactly selected recipes transfer to four unseen benchmarks, and controlled analyses show that recovery depends on both where and what parameters are restored. Interventional analysis further links the selected layers to temporal reasoning sensitivity, and qualitative attribution reveals corresponding changes in temporal evidence use. More broadly, multimodal adaptation need not permanently erase prior competence: when the original capability remains encoded in a paired checkpoint, targeted parameter recovery can restore it while retaining what adaptation added.

AI USE STATEMENT

Generative AI tools were used for language polishing and limited feedback on the presentation and organization of the manuscript. All research conceptualization, methodological decisions, experimental execution, validation, analysis, and final scientific judgments were made by the authors. Separately, pretrained language and multimodal models were used as components of the experimental methodology, including LLM-based evaluation following official benchmark protocols, as described in the paper. The authors reviewed all AI-assisted work and take full responsibility for the final content.

ETHICS STATEMENT

This work uses existing models and benchmark datasets and does not involve human-subject studies or new data collection. We do not identify additional ethical concerns beyond those associated with the underlying models and datasets.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings, model pairings, evaluation metrics, baseline configurations, implementation details, and search procedures in the appendix, together with complete results for all evaluated benchmarks. Code and data are available at the anonymous repository linked in the paper: https://anonymous.4open.science/r/MERIT-ICLR27.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7(2):195–204, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. Rextime: A benchmark suite for reasoning-across-time in videos. *arXiv preprint arXiv:2406.19392*, 2024.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring reason to vision: Understanding perception and reasoning through model merging. *arXiv preprint arXiv:2505.05464*, 2025.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Lost in time: A new temporal benchmark for videollms. *arXiv preprint arXiv:2410.07752*, 2024.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's mergekit: A toolkit for merging large language models. *arXiv preprint arXiv:2403.13257*, 2024.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024.

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. Temporal reasoning transfer from text to video. *arXiv preprint arXiv:2410.06166*, 2024a.

Shengzhi Li, Rongyu Lin, and Shichao Pei. Multi-modal preference alignment remedies degradation of visual instruction tuning on language models. *arXiv preprint arXiv:2402.10884*, 2024b.

Zhaohe Liao, Jiangtong Li, Siyu Sun, Qingyang Liu, Fengshun Xiao, Tianjiao Li, Qiang Zhang, Guang Chen, Li Niu, Changjun Jiang, et al. Divide and conquer: Exploring language-centric tree reasoning for video question-answering. In *Forty-second International Conference on Machine Learning*, 2025.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13235–13245, 2024.

Ming Nie, Dan Ding, Chunwei Wang, Yuanfan Guo, Jianhua Han, Hang Xu, and Li Zhang. Slowfocus: Enhancing fine-grained temporal understanding in video LLM. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Daiqing Qi, Dongliang Guo, Hanzhang Yuan, Handong Zhao, Mengxuan Hu, Lehan Yang, and Sheng Li. Improve temporal reasoning in multimodal large language models via video contrastive decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Haiyi Qiu, Minghe Gao, Long Qian, Kaihang Pan, Qifan Yu, Juncheng Li, Wenjie Wang, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. Step: Enhancing video-llms' compositional reasoning by spatio-temporal graph-guided self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3284–3294, 2025.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint arXiv:2410.23266*, 2024.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *The Twelfth International Conference on Learning Representations*, 2024.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David Blei. Hypothesis testing the circuit hypothesis in LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-llm reasoning via agent-of-thoughts distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8523–8533, 2025.

Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. An empirical study of multimodal model merging. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1563–1575, 2023.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14820–14835, 2023.

Haonan Wang, Weida Liang, Zihang Fu, Nie Zheng, Yifan Zhang, Yao Tong, Tongyao Zhu, Hao Jiang, Chuang Li, Jiaying Wu, and Kenji Kawaguchi. From harm to help: Turning reasoning in-context demos into assets for reasoning lms. *arXiv preprint arXiv:2509.23196*, 2025a.

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, 2025b.

Yuqing Wang and Yun Zhao. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024a.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025b.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations*, 2024b.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Zeping Yu and Sophia Ananiadou. Locate-then-merge: Neuron-level parameter fusion for mitigating catastrophic forgetting in multimodal LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7065–7078, 2025.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024a.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b.

Xinyu Zhang, Yuxuan Dong, Lingling Zhang, Chengyou Jia, ZhuoHang Dang, Yixing Yao, Yaqiang Wu, Basura Fernando, and Jun Liu. Beyond layer-wise merging: Chain-of-merging for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24279–24289, 2026.

Didi Zhu, Yibing Song, Tao Shen, Ziyu Zhao, Jinluan Yang, Min Zhang, and Chao Wu. REMEDY: Recipe merging dynamics in large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.

Yuchen Zou, Leyan Wang, Qinghong Zhao, Yipan Wei, Yapeng Li, and Bo Du. Injecting reasoning into vision-language models via weight-decomposed merging. *Neural Networks*, pp. 109364, 2026.

List of Appendices

A Limitations and Future Work **15**

B Text-Only Recovery at the Language-Backbone Level **15**

C Experimental Details **16**

C.1 Model Pairings 16

C.2 Evaluation Tasks and Metrics 16

C.3 Baseline Construction 16

C.4 Implementation and Evaluation 17

D Additional Analyses of MERIT **17**

D.1 Search Effectiveness and Recipe Stability 17

D.2 Intervention-Module Ablations 18

D.3 Search Cost and Deployment Efficiency 19

D.4 Recovery Extends to a Larger VLM 19

D.5 Additional Temporal-Grounding Case Studies 19

E Interpolation-Controlled Merging Baselines **23**

F Benchmark Details and Complete Results **24**

F.1 Benchmark Statistics and Task Formats 24

F.2 Complete Results by Benchmark 25

A Limitations and Future Work

By framing adaptation-induced capability loss as a selective recovery problem, our study opens several directions beyond the current formulation.

First, **MERIT currently relies on benchmark-defined signals for temporal perception and temporal reasoning.** This controlled separation makes recovery measurable, but assumes access to clean target and preservation signals. A natural next step is to replace them with weaker supervision, synthetic diagnostics, preference signals, or model-based evaluation. This would extend selective recovery to settings where degraded and protected capabilities are only partially observed.

Second, **the current parameterization deliberately favors a compact, tractable repair space.** MERIT searches discrete layer-dominance patterns while selecting interpolation strength through an outer grid. Our finding that recovery depends on both where and what parameters are restored suggests a richer space that could jointly optimize layer selection, interpolation strength, intervention modules, or finer-grained structures. More sample-efficient black-box optimization could make such adaptive repair practical.

Finally, **model-specific recovery recipes offer a new lens on how capabilities are altered by adaptation.** Systematic comparisons across architectures and adaptation trajectories could reveal recurring intervention patterns, transferable search priors, or reusable rules for capability repair. Extending the same framework beyond temporal reasoning would further test which pretrained capabilities remain recoverable after multimodal adaptation. More broadly, this opens a path toward understanding when capability loss is reversible, where recoverable competence remains encoded, and how such recovery can be automated.

B Text-Only Recovery at the Language-Backbone Level

To test whether MERIT repairs the language-side capability identified in Figure 1, we evaluate temporal reasoning without visual input. Comparing the original LLM, the adapted language backbone, and MERIT isolates whether the recovery observed on video tasks extends to the temporal competence degraded within the language component itself.

Benchmark and Tasks. We evaluate on TempReason (Tan et al., 2023), focusing on its two higher-level tasks: L2 time–event reasoning, which identifies the event associated with a queried time, and L3 event–event reasoning, which infers temporal relations between events. The evaluation contains 5,397 L2 and 4,426 L3 examples.

Evaluation Protocol. Within each model family, the original LLM, adapted language backbone (VLM Text-Only), and MERIT use the same prompt, with each model formatted using its native tokenizer and chat template:

```
Answer the question using only the temporal facts in the context.
Return only a short answer, without explanation.
Context: {context}
Question: {question}
Answer:
```

Decoding is deterministic with temperature 0 and stops at the first newline. We report lenient Exact Match (EM, %): after normalization, a prediction is correct if it exactly matches or contains a reference answer, accommodating minor wrappers produced by chat models.

Backbone-Level Recovery. Figure 5 shows a consistent pattern across all three model families: multimodal adaptation weakens text-only TR, and MERIT recovers part or all of the lost performance. For LongVA-7B, MERIT improves the adapted backbone by 6.20%, 4.64%, and 5.57% on time–event, event–event, and overall EM, respectively, surpassing the original LLM on all three metrics. InternVL3-8B exhibits the largest degradation, especially on event–event reasoning, where EM falls from 16.40 to 13.24; MERIT raises it to 14.63 and improves overall EM by 9.63%. Qwen3-VL-4B follows the same pattern, with overall EM decreasing from 18.18 to 17.15 after adaptation and

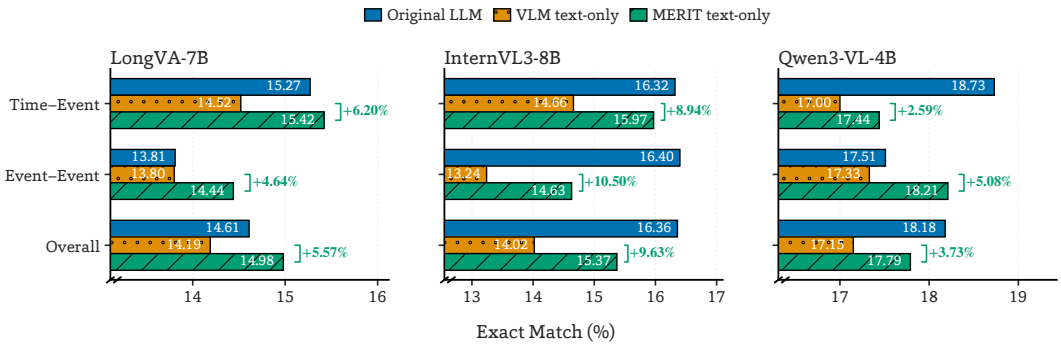

Figure 5: **MERIT recovers temporal reasoning weakened within the adapted language backbone.** Across all three model families, multimodal adaptation reduces text-only TEMPREASON performance, while MERIT restores part or all of the lost capability. Percentages beside the brackets show relative gains over the adapted backbone.

recovering to 17.79 with MERIT. These results close the loop between diagnosis and recovery: the intervention that improves video reasoning also restores temporal competence within the adapted language backbone.

C EXPERIMENTAL DETAILS

C.1 MODEL PAIRINGS

We pair each VLM with the pre-adaptation language backbone specified by its original release: LongVA-7B with Qwen2-7B-Instruct (Yang et al., 2024a), InternVL3-8B with Qwen2.5-7B-Instruct (Yang et al., 2025b), and Qwen3-VL-4B-Instruct with Qwen3-4B-Instruct-2507 (Yang et al., 2025a). LongVA-7B and InternVL3-8B contain 28 self-attention layers, while Qwen3-VL-4B contains 36.

C.2 EVALUATION TASKS AND METRICS

We report each benchmark's metric most closely aligned with temporal reasoning: Temporal Reasoning for Video-MME and MMBench-Video, Relation for LongVideoBench, and Reasoning for LVBench. For Video-Holmes, whose questions center on multi-step inference, we report overall accuracy. Dataset statistics and task formats are summarized in § F.1.

C.3 BASELINE CONSTRUCTION

FULL-BACKBONE interpolates the complete language backbone, while ALL-ATTENTION interpolates every self-attention layer, following the parameter-level merging setting of Chen et al. (2025). RANDOM-K uniformly samples a binary layer pattern with exactly k self-attention layers assigned to the $\hat{u}_\ell = 1$ branch of Eq. 4, where k matches the number of layers assigned $\hat{u}^*_\ell = 1$ in the corresponding MERIT recipe; the remaining layers follow the $\hat{u}_\ell = 0$ branch. We report mean and standard deviation over five runs. For the fixed-α RANDOM-K baseline, we use the interpolation coefficient selected by MERIT for the corresponding model. For α-tuned baselines, the interpolation coefficient is selected on the same TP/TR diagnostic set used by MERIT; for RANDOM-K, α is tuned independently for each sampled layer pattern.

LORA and full-parameter fine-tuning (FFT) use the same 232 diagnostic examples for training. LORA applies rank-16 adapters to the language model with a learning rate of 4×10^{-5}, while FFT updates the full language backbone with a learning rate of 5×10^{-6}. Both use bfloat16, cosine scheduling, weight decay of 0.05, and a warmup ratio of 0.03. Training budgets and GPU usage are reported in Table 7.

Table 3: **MERIT discovers distinct recovery recipes across model families.** We report the selected interpolation weight α, number of LLM-dominant self-attention layers ($\hat{u}_\ell^* = 1$), and their 0-based indices.

VLM	α	#Layers	$\hat{u}_\ell^* = 1$ self-attention layer indices
LongVA-7B	1.0	11	1, 5, 7–8, 11, 13–14, 16, 19–20, 23
InternVL3-8B	0.9	9	4, 7, 12, 14, 16–18, 24, 26
Qwen3-VL-4B	1.0	16	0, 4, 8, 10, 13, 17–18, 21, 23, 26, 28–32, 35

C.4 Implementation and Evaluation

Our implementation builds on `mergekit-evolve` (Goddard et al., 2024). For each fixed α, CMA-ES uses a population size of $\lambda_{\text{pop}} = 4 + \lfloor 3 \ln L \rfloor$ and evaluates at most 1,600 candidates under $\mathcal{F}$. We cache objective values for repeated discrete recipes and visual-processor outputs, yielding a 2–3$\times$ speedup. We set the perceptual-degradation weight to $\lambda = 55/177$, matching the sizes of the TP and TR diagnostic subsets so that a one-example decrease in TP accuracy carries the same weight as a one-example increase in TR accuracy.

We evaluate Video-MME, LongVideoBench, and LVBench with `lmms-eval` (Zhang et al., 2025), and MMBench-Video and Video-Holmes with VLMEvalKit (Duan et al., 2024). Following the official benchmark implementations (Fang et al., 2024; Cheng et al., 2025), MMBench-Video and Video-Holmes use GPT-4-Turbo and GPT-4-0125 (Achiam et al., 2023) as judges. Decoding is deterministic with temperature 0. We use at most 128 frames for LongVA and 64 frames for InternVL3 and Qwen3-VL; for Qwen3-VL, we set `min_pixels`=8,192 and `max_pixels`=786,432. MMBench-Video reports the mean judge score on the $\{0, 1, 2, 3\}$ scale; all other benchmarks report accuracy.

D Additional Analyses of MERIT

This section provides complementary analyses of MERIT beyond the main results. We examine the effectiveness and stability of evolutionary recipe search (§ D.1), the choice of intervention module (§ D.2), search cost and deployment efficiency (§ D.3), scaling to a larger VLM (§ D.4), and additional temporal-grounding case studies (§ D.5).

D.1 Search Effectiveness and Recipe Stability

Recovered recipes are model-specific. Table 3 reports the best recipe found for each VLM. The selected configurations differ in interpolation strength, number of LLM-dominant layers, and intervention locations. No single layer pattern therefore emerges across model families, motivating model-specific recipe search.

CMA-ES improves over matched-budget random search. We compare CMA-ES with uniform random search under the same α, layer-recipe space, perception-constrained objective $\mathcal{F}(g)$, and 1,600-candidate budget. For each VLM, both methods use the α selected by MERIT in Table 3; only the candidate-generation strategy differs. We run three independent trials for each method.

Figure 6 shows that CMA-ES reaches a higher mean final objective than random search on all three VLMs. Its three trials also recover the same final objective and recipe for each model despite different intermediate trajectories. This comparison complements the α-tuned Random-K baseline, which separately controls for interpolation-strength selection across random layer sets.

The search advantage transfers downstream. For InternVL3-8B, we evaluate the best recipe from each random-search run on Video-MME, LVBench, and Video-Holmes using the same $\alpha = 0.9$ as MERIT. As shown in Table 4, random search improves Video-MME TR over Base but remains below MERIT on both the diagnostic task and unseen benchmarks. The advantage of CMA-ES therefore extends beyond optimization of the diagnostic objective.

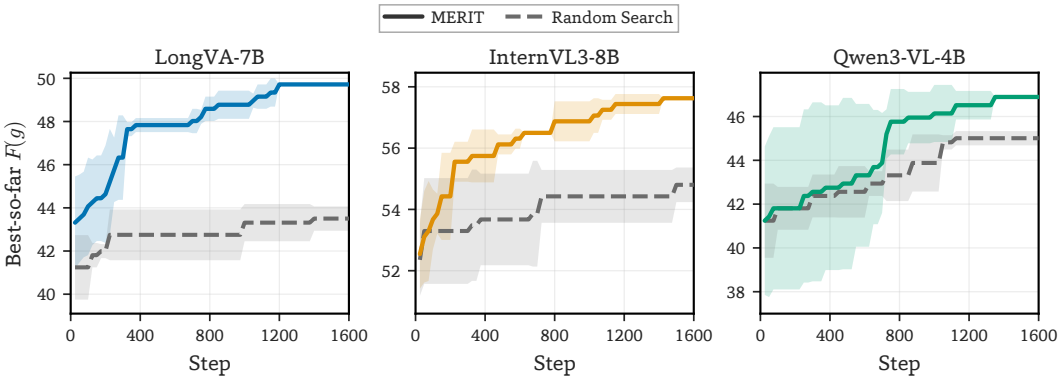

Figure 6: **CMA-ES outperforms matched-budget uniform random search.** Best-so-far perception-constrained objective $\mathcal{F}(g)$ over 1,600 candidate evaluations under the same α, recipe space, objective, and budget. Curves show the mean over three runs; shaded regions indicate ± 1 standard deviation.

Table 4: **Downstream comparison of MERIT and matched-budget random search on InternVL3-8B.** Both use $\alpha = 0.9$. Random-search values report mean $\pm$ standard deviation over the best recipes from three independent runs; the three MERIT runs recover the same final recipe and are therefore reported once. All values are accuracy (%).

Method	Video-MME		LVBench		Video-Holmes
	TR	Overall	TR	Overall	Overall
BASE	52.0	66.5	51.2	44.8	36.5
RANDOM SEARCH	54.8±0.6	66.4±0.4	49.7±1.8	43.8±0.6	36.6±0.6
MERIT	**57.6**	**67.9**	**53.2**	**45.0**	**37.2**

D.2 INTERVENTION-MODULE ABLATIONS

Interpolation weight under a fixed layer pattern. To test whether the recovery pattern found for self-attention transfers to other parameter groups, we fix the binary layer pattern $\hat{\mathbf{u}}^*$ discovered by MERIT for InternVL3-8B and apply the directional interpolation in Eq. 4 to MLP-only, attention-plus-MLP, and full-block parameters. For each variant, only α is tuned on the same 232-example TP/TR diagnostic set.

Here, the complement symmetry used by the original MERIT search no longer removes coefficients below 0.5, because the layer pattern is held fixed rather than complemented jointly with α. We therefore sweep $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$. Table 5 reports the complete results.

The alternative modules favor substantially smaller interpolation weights: MLP-only selects $\alpha = 0.1$, while attention-plus-MLP and full-block merging select $\alpha = 0.3$. This shift indicates that the interpolation direction effective for self-attention does not transfer directly to MLP-related parameters.

Self-attention provides the strongest transferable recovery. Using the selected α for each variant, we evaluate the resulting models on Video-MME, LVBench, and Video-Holmes. Table 6 compares them with BASE and the original attention-only MERIT configuration.

Attention-only merging provides the strongest and most consistent gains. It raises Video-MME TR from 52.0 to 57.6, LVBench Reasoning from 51.2 to 53.2, and Video-Holmes accuracy from 36.5 to 37.2. MLP-only merging provides no Video-MME TR gain and reduces LVBench Reasoning to 49.8, while attention-plus-MLP and full-block interventions remain below attention-only MERIT on held-out evaluations. Thus, the selected layer pattern alone is insufficient: effective recovery also depends on which parameters are modified within those layers.

Table 5: **Full interpolation-weight grid for the intervention-module ablation.** Each entry reports the TP/TR search objective $\mathcal{F}(g)$ on the 232-example diagnostic set. The binary layer pattern $\hat{\mathbf{u}}^*$ discovered by MERIT is fixed across all variants, while the merged parameter group and interpolation coefficient α are varied. Bold indicates the selected α for each variant.

Model	Variant	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
InternVL3-8B	MLP ONLY	–	**52.0**	51.4	51.4	51.4	50.3	50.8	49.7	46.3	46.9	45.8
	ATTENTION + MLP	–	52.5	54.2	**54.8**	54.2	53.7	52.5	52.0	50.3	49.7	46.9
	FULL BLOCK	–	51.4	53.7	**54.8**	54.2	54.2	53.1	52.0	50.8	49.7	47.5

Table 6: **Ablation over the intervention module on InternVL3-8B.** All non-base variants use the same fixed binary layer pattern $\hat{\mathbf{u}}^*$ discovered by MERIT, with α tuned on the TP/TR diagnostic set. We report accuracy (%) on the full Video-MME, LVBench, and Video-Holmes benchmarks. Best results are shown in bold.

Variant	α	Video-MME		LVBench		Video-Holmes
		TR	Overall	TR	Overall	Overall
BASE	–	52.0	66.5	51.2	44.8	36.5
ATTENTION ONLY (MERIT)	0.9	**57.6**	**67.5**	**53.2**	**45.0**	**37.2**
MLP ONLY	0.1	52.0	66.5	49.8	44.6	36.5
ATTENTION + MLP	0.3	54.8	66.7	51.2	44.6	36.7
FULL BLOCK	0.3	54.8	66.7	50.7	44.7	36.7

D.3 SEARCH COST AND DEPLOYMENT EFFICIENCY

Table 7 reports computational cost on InternVL3-8B using NVIDIA A100-SXM4-80GB GPUs and the same 232-example TP/TR diagnostic set. FFT and LoRA use these examples for parameter optimization, whereas MERIT uses them only as black-box signals for recipe selection.

Each fixed-α MERIT search evaluates at most 1,600 candidates and takes approximately 1.7 hours using 4 GPUs and 16 parallel workers. As described in § 3.3, MERIT searches five interpolation strengths independently, while $\alpha = 0.5$ requires only one evaluation. The complete grid therefore contains at most 8,001 candidate evaluations ($5 \times 1{,}600$ search evaluations plus one evaluation at $\alpha = 0.5$). Executed sequentially with the same 4-GPU allocation, the five searches require approximately 8.5 hours plus one single-candidate evaluation; they can instead be parallelized across α.

All search cost is incurred offline. MERIT performs no gradient updates, introduces no trainable parameters, and compiles the selected recipe directly into the checkpoint. The resulting model therefore retains the architecture and inference cost of the original VLM.

D.4 RECOVERY EXTENDS TO A LARGER VLM

To test whether the recovery pattern extends beyond the 4B–8B models used in the main evaluation, we apply MERIT to InternVL3-14B. As shown in Table 8, MERIT improves Video-MME TR from 59.3 to 66.7, LVBench Reasoning from 48.8 to 50.7, and Video-Holmes accuracy from 37.3 to 38.0, while also improving overall Video-MME and LVBench performance. It outperforms full-backbone, all-attention, and α-tuned RANDOM-K merging on every reported metric. The same selective-recovery pattern therefore persists at a larger model scale.

D.5 ADDITIONAL TEMPORAL-GROUNDING CASE STUDIES

We present three additional Video-Holmes cases using the same frame-level attribution and reasoning-trace analysis as in the main text. Together, they illustrate how recovery can alter temporal grounding both when it corrects an error and when the final answer is unchanged.

Table 7: **Runtime and resource cost on InternVL3-8B.** All measurements use NVIDIA A100-SXM4-80GB GPUs and the same 232-example TP/TR diagnostic set. The MERIT row reports one fixed-α search; full-grid cost is detailed in § D.3.

Method	Data Usage	Resources	Optimization Budget	Gradient Updates	Trainable Parameters	Wall-Clock Time
FFT	232 training examples	8 GPUs	1 epoch / 29 steps	Yes	7.61B	∼0.5 h
LoRA	232 training examples	4 GPUs	3 epochs / 174 steps	Yes	40.37M	∼1.6 h
One merging candidate	232 validation examples	1 GPU / 1 worker	1 candidate	No	0	∼2 min
MERIT (per α)	232 search examples	4 GPUs / 16 workers	1,600 candidates	No	0	∼1.7 h

Table 8: **Larger-scale validation on InternVL3-14B.** MERIT outperforms full-backbone, all-attention, and α-tuned RANDOM-K interpolation across all reported reasoning-aligned and overall metrics. Best performance in each evaluation column is shown in bold. † denotes α-tuned RANDOM-K.

Model	Variant	Best α	Video-MME		LVBench		Video-Holmes
			TR	Overall	TR	Overall	Overall
InternVL3-14B	BASE	N/A	59.3	70.0	48.8	45.4	37.3
	FULL-BACKBONE	0.1	60.5	69.9	49.3	45.9	37.4
	ALL-ATTENTION	0.1	61.0	70.2	49.3	46.2	36.6
	RANDOM-K†	0.8, 0.6, 0.7, 0.0, 1.0	59.8±1.2	70.0±0.2	47.9±1.4	45.9±0.4	37.0±0.2
	MERIT	1.0	**66.7**	**70.6**	**50.7**	**46.9**	**38.0**

In Figure 7, both models answer correctly, but MERIT links the apparent theft, subsequent pursuit, and later awakening into a longer event chain supporting the dream interpretation. In Figure 8, BASE anchors on the murderer's entry through an unlocked door, whereas MERIT integrates the staged scare, repeated screams, and ignored warning to recover the correct causal explanation. In Figure 9, both models again answer correctly, but MERIT connects repeated wake-up and gun-firing sequences to infer a recurring temporal pattern. These examples show that recovery can change the temporal evidence structure underlying a prediction even when the final answer itself does not change.



Figure 7: **Case study from Video-Holmes (Question ID: 1040).** Both models answer correctly, but MERIT grounds its prediction in a longer event chain linking the apparent theft, the pursuit, and the later awakening. This temporal structure supports the inference that the preceding events were a dream, while the base model remains more locally grounded.

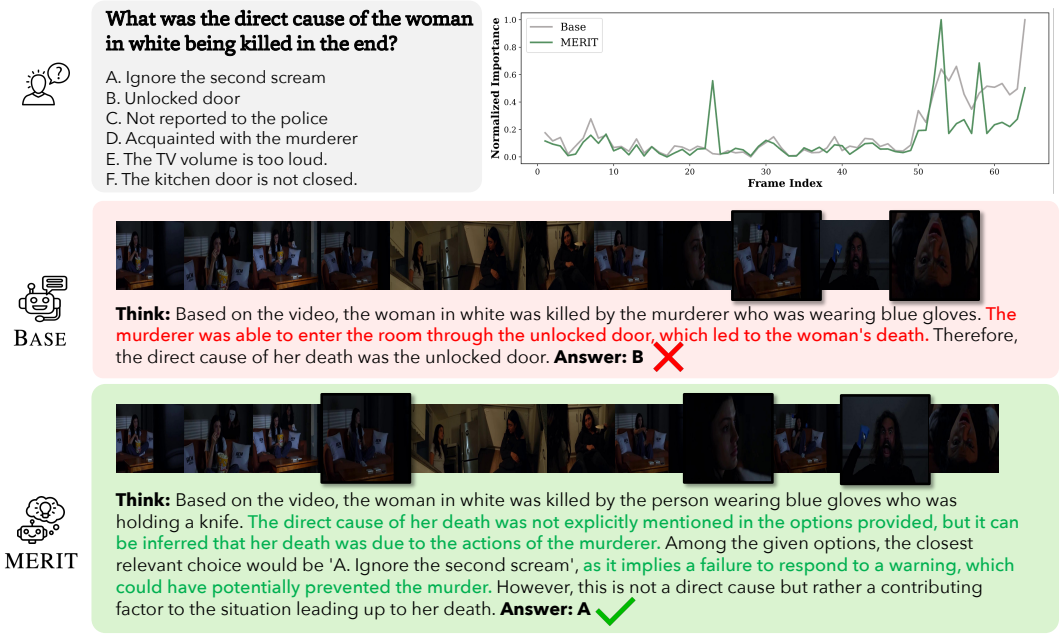

Figure 8: **Case study from Video-Holmes (Question ID: 565).** The base model anchors on the murderer's entry through the unlocked door and incorrectly treats it as the direct cause of death. MERIT instead links the staged scare, the first non-lethal encounter, and the ignored second scream, correctly identifying the missed warning as the decisive causal factor.

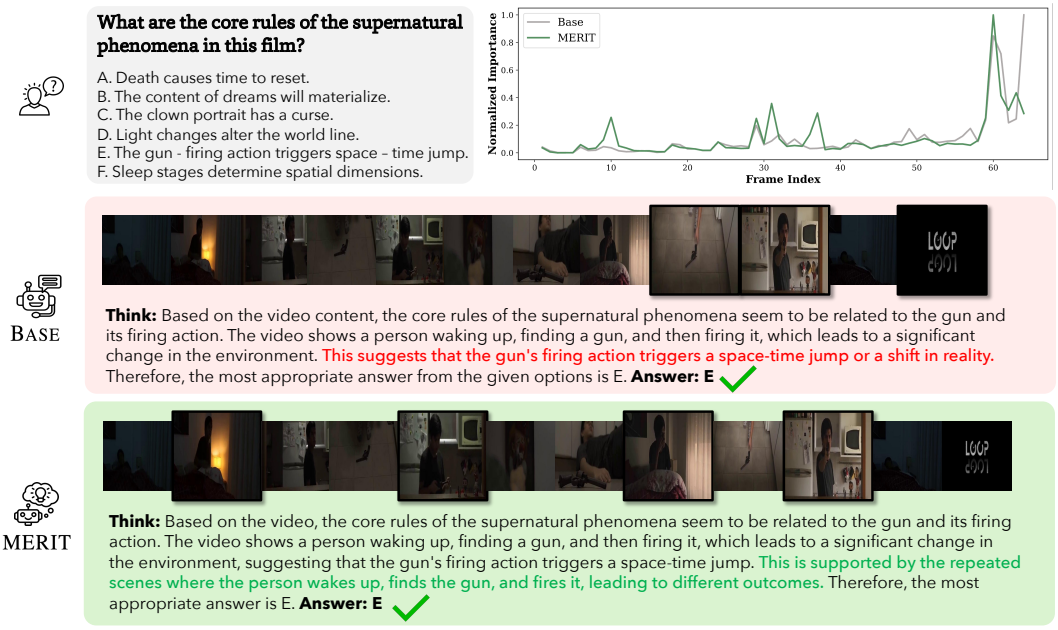

Figure 9: **Case study from Video-Holmes (Question ID: 527).** Both models select the correct answer, but MERIT links repeated wake-up and gun-firing sequences to infer a recurring pattern in which each gunshot triggers a shift in reality. The base model relies primarily on a single observed transition.

Table 9: **Full alpha-grid results on the TP/TR diagnostic set.** Each entry reports the search objective $\mathcal{F}(g)$ for the corresponding interpolation coefficient. For each RANDOM-K run, the binary layer pattern is fixed while α is swept across the full grid. Bold marks the selected α; objective ties are broken in favor of the higher TP score to avoid unnecessary perceptual degradation for TR improvement.

Model	Variant	Run	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
LongVA-7B	FULL-BACKBONE	–	–	38.4	**40.1**	39.5	37.3	33.9	31.1	29.9	19.8	4.5	-1.1
	ALL-ATTENTION	–	–	**39.5**	39.0	39.5	39.0	39.5	37.9	39.0	37.3	35.6	32.8
	RANDOM-K ATTENTION	1	36.2	39.0	39.0	38.4	39.0	39.5	39.5	39.5	39.5	39.0	**40.1**
		2	37.9	37.3	37.3	39.0	39.0	39.5	39.0	39.5	39.5	**41.2**	38.4
		3	38.4	39.0	39.5	39.0	39.5	**39.5**	39.5	39.0	38.4	36.7	37.9
		4	39.0	41.2	**41.8**	40.1	39.5	39.5	39.0	37.9	36.2	36.2	35.6
		5	35.0	36.2	35.6	37.9	37.3	39.5	39.5	40.1	40.1	**40.7**	38.4
InternVL3-8B	FULL-BACKBONE	–	–	53.1	**53.1**	50.3	49.2	48.6	42.9	40.1	36.7	32.2	27.7
	ALL-ATTENTION	–	–	**54.2**	53.7	54.2	54.2	51.4	51.4	50.8	49.2	49.2	50.3
	RANDOM-K ATTENTION	1	52.0	52.0	51.4	50.8	51.4	51.4	**52.0**	50.8	51.4	51.4	50.3
		2	49.7	51.4	52.5	**53.7**	52.5	51.4	52.0	52.5	51.4	50.8	50.3
		3	52.5	51.4	52.5	**52.5**	52.0	51.4	52.5	52.0	52.5	51.4	52.0
		4	52.0	50.3	52.0	52.0	52.5	51.4	**53.1**	52.5	52.0	51.4	49.2
		5	48.0	49.7	51.4	51.4	52.5	51.4	**54.2**	53.7	52.0	53.1	51.4
Qwen3-VL-4B	FULL-BACKBONE	–	–	42.9	44.1	**45.2**	44.1	39.5	35.6	32.2	26.0	24.3	13.6
	ALL-ATTENTION	–	–	**44.1**	42.9	42.9	43.5	41.2	39.5	40.1	39.0	35.0	33.3
	RANDOM-K ATTENTION	1	39.0	38.4	41.2	41.8	42.4	41.2	**45.2**	44.6	43.5	42.9	42.9
		2	40.7	42.4	**43.5**	42.9	42.4	41.2	41.2	41.2	39.5	39.5	37.3
		3	40.7	42.9	**45.8**	44.6	42.4	41.2	41.2	41.8	41.8	42.4	38.4
		4	43.5	43.5	43.5	42.9	**44.1**	41.2	40.7	40.7	42.4	39.0	37.3
		5	38.4	39.0	38.4	38.4	42.4	41.2	**42.9**	42.4	41.2	37.9	42.4

E INTERPOLATION-CONTROLLED MERGING BASELINES

This section details how we construct the merging baselines used to isolate the effects of the interpolation coefficient and layer selection.

Controlling for interpolation strength. We independently tune α for FULL-BACKBONE, ALL-ATTENTION, and each RANDOM-K baseline on the same 232-example TP/TR diagnostic set used by MERIT. FULL-BACKBONE and ALL-ATTENTION use $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$, with $\alpha = 0$ omitted because it reduces to the unmodified VLM. For RANDOM-K, each run fixes a randomly sampled binary layer pattern with exactly k layers assigned to the $\hat{u}_\ell = 1$ branch of Eq. 4, where k matches the number of such layers in the corresponding MERIT recipe. Because this layer pattern is fixed, the complement symmetry used by MERIT does not make α and $1 - \alpha$ redundant; complementing the pattern would generally also change its cardinality from k to $L - k$. We therefore sweep the full grid $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$ for RANDOM-K. Table 9 reports the complete sweep.

The preferred interpolation coefficient depends on both intervention granularity and layer identity. FULL-BACKBONE generally favors small α and degrades as more weight is placed on the pre-adaptation checkpoint, while ALL-ATTENTION selects $\alpha = 0.1$ for all three models. For RANDOM-K, the preferred coefficient varies substantially across fixed random layer patterns, showing that interpolation direction and layer identity are coupled.

Alpha-tuned RANDOM-K controls for layer identity. Table 10 reports the five random layer patterns used for each backbone. Every run assigns exactly k self-attention layers to the $\hat{u}_\ell = 1$ branch, with k matching the corresponding MERIT recipe, while the remaining layers follow the

Table 10: **Alpha-tuned random-k baseline recipes.** For each model, we report five independent random binary layer patterns, each assigning exactly k self-attention layers to the $\hat{u}_\ell = 1$ branch of Eq. 4, with k matching the number of layers assigned $\hat{u}^*_\ell = 1$ in the corresponding MERIT recipe. The interpolation coefficient α is tuned independently for every run on the TP/TR diagnostic set. Because α may be below 0.5, the $\hat{u}_\ell = 1$ branch is not necessarily LLM-dominant. Layer indices are 0-based.

VLM	α	#Layers	Run	$\hat{u}_\ell = 1$ self-attention layer indices
LongVA-7B	1.0	11	1	5, 7–8, 10–11, 17, 20–22, 26–27
	0.9		2	0, 3–6, 8–9, 14–15, 19, 27
	0.5		3	1, 4, 7, 11–12, 15, 20–21, 23, 26–27
	0.2		4	1, 6–7, 15, 17–18, 20–22, 25, 27
	0.9		5	5, 8–9, 12–15, 21, 24–25, 27
InternVL3-8B	0.6	9	1	5, 8, 10, 12–13, 16, 23–24, 27
	0.3		2	0–1, 3–4, 6, 10, 12, 16, 23
	0.3		3	0, 2–3, 9, 13, 15–16, 18, 26
	0.6		4	0–1, 4, 12, 16, 19, 24, 26–27
	0.6		5	1, 6, 8, 10, 18–19, 22, 26–27
Qwen3-VL-4B	0.6	16	1	2–3, 6–7, 12, 14, 21–25, 28–31, 33
	0.2		2	0, 3–7, 9–11, 15, 17, 22, 24, 30, 33, 35
	0.2		3	1–4, 7–8, 10, 14–15, 17–18, 20–23, 25
	0.4		4	0–2, 7, 11, 14–15, 18–19, 21, 25–28, 32, 35
	0.6		5	0, 3–4, 6, 8–10, 14, 17–18, 20–21, 24, 30–31, 35

$\hat{u}_\ell = 0$ branch. For each fixed pattern, α is tuned independently using the grid in Table 9. These configurations form the α-tuned RANDOM-K baseline reported in the main results.

Independent tuning prevents interpolation coefficient selection from favoring MERIT. Since the random patterns match the cardinality of the MERIT pattern but not its layer identities, their remaining performance gap supports the contribution of the discovered layer configuration.

F BENCHMARK DETAILS AND COMPLETE RESULTS

This section summarizes the evaluated benchmarks and reports the complete per-benchmark results underlying the main-text comparisons.

F.1 BENCHMARK STATISTICS AND TASK FORMATS

- **Video-MME** (Fu et al., 2025): A broad VideoQA benchmark covering diverse video-understanding capabilities, including long-context temporal reasoning. It contains 900 videos and 2,700 multiple-choice questions.
- **LongVideoBench** (Wu et al., 2024): A long-context video–language benchmark requiring retrieval and reasoning over temporally localized segments. Its validation set contains 752 videos and 1,337 multiple-choice questions.
- **LVBench** (Wang et al., 2025b): A long-video understanding benchmark covering temporal grounding, reasoning, and related capabilities. It contains 103 videos longer than 30 minutes and 1,549 question–answer pairs with temporal annotations.
- **MMBench-Video** (Fang et al., 2024): A VideoQA benchmark for fine-grained multi-shot understanding and temporal reasoning, with 609 videos and 1,998 open-ended question–answer pairs evaluated using an LLM-based scoring protocol.
- **Video-Holmes** (Cheng et al., 2025): A reasoning-centric benchmark built from suspense videos, designed to test multi-step inference through cross-segment clue aggregation. It contains 270 videos and 1,837 multiple-choice questions.

F.2 Complete Results by Benchmark

We report the complete metrics underlying the summarized main-text results: Table 11 for Video-MME, Table 12 for LongVideoBench, Table 13 for LVBench, Table 14 for MMBench-Video, and Table 15 for Video-Holmes.

Table 11: **Full results (Accuracy %) on Video-MME.** TP: Temporal Perception; SP: Spatial Perception; AP: Attribute Perception; ActRec: Action Recognition; ObjRec: Object Recognition; OCR: OCR Problems; Cnt: Counting Problem; TR: Temporal Reasoning; SR: Spatial Reasoning; AR: Action Reasoning; OR: Object Reasoning; IS: Information Synopsis.

Model	Variant	Run	Overall	TP	SP	AP	ActRec	ObjRec	OCR	Cnt	TR	SR	AR	OR	IS
LongVA-7B	BASE	–	54.1	60.0	64.8	67.1	53.4	60.2	55.4	30.6	40.1	62.5	48.1	52.4	69.7
	LORA	–	55.0	72.7	66.7	67.6	54.3	58.8	56.1	31.0	53.1	67.9	48.4	50.0	69.0
	FFT	–	54.3	81.8	63.0	65.8	50.8	54.8	55.4	32.1	64.4	67.9	44.6	50.2	67.5
	FULL-BACKBONE	–	54.1	60.0	63.0	66.2	51.8	59.3	55.4	31.0	40.1	66.1	49.1	52.4	70.9
	ALL-ATTENTION	–	54.3	61.8	64.8	65.8	53.4	59.3	56.1	31.0	39.5	67.9	48.8	52.6	70.6
	RANDOM-K	1	54.2	60.0	64.8	67.6	52.1	59.0	55.4	30.6	40.1	66.1	49.5	52.0	71.2
		2	53.3	60.0	57.4	63.1	50.8	60.2	56.1	28.7	38.4	62.5	47.7	52.6	71.2
		3	53.6	56.4	66.7	64.0	50.8	59.6	58.3	32.5	39.0	66.1	48.4	51.5	69.0
		4	53.2	60.0	64.8	63.1	51.1	59.3	54.7	29.1	35.6	64.3	50.2	51.3	70.9
		5	54.4	58.2	63.0	69.8	53.4	59.6	54.0	31.7	39.0	67.9	50.5	50.9	70.9
	RANDOM-K (α-tuned)	1	54.2	60.0	64.8	67.6	52.1	59.0	55.4	30.6	40.1	66.1	49.5	52.0	71.2
		2	53.8	60.0	59.3	64.0	50.2	61.0	56.8	31.0	41.2	64.3	47.4	52.2	71.2
		3	53.7	60.0	63.0	65.3	50.5	60.5	55.4	29.5	39.5	64.3	48.8	52.4	70.3
		4	53.4	60.0	59.3	65.8	49.5	60.7	56.8	26.1	41.8	69.6	48.1	51.5	70.9
		5	54.7	58.2	63.0	69.8	53.7	58.8	55.4	31.3	41.2	67.9	50.2	51.8	70.9
	MERIT	–	55.1	61.8	64.8	68.0	52.7	59.9	57.6	29.9	49.7	69.6	48.1	53.3	69.3
InternVL3-8B	BASE	–	66.5	74.5	64.8	78.8	64.2	72.6	72.7	46.6	52.0	83.9	61.4	64.1	79.3
	LORA	–	68.0	83.6	64.8	78.8	65.8	72.3	71.9	45.9	71.2	82.1	61.1	63.9	80.2
	FFT	–	67.0	85.5	63.0	77.0	65.2	72.0	69.1	47.0	71.2	76.8	60.4	61.9	78.6
	FULL-BACKBONE	–	66.6	76.4	64.8	77.0	66.8	70.9	71.9	47.0	53.1	82.1	61.1	63.7	80.8
	ALL-ATTENTION	–	66.6	76.4	63.0	77.9	64.2	72.6	71.2	48.5	54.2	83.9	61.1	63.7	79.6
	RANDOM-K	1	66.6	80.0	66.7	77.9	66.1	72.9	71.9	46.6	51.4	83.9	58.9	63.7	80.5
		2	67.0	78.2	64.8	76.1	66.1	73.4	76.3	50.4	50.8	80.4	60.0	64.1	79.9
		3	66.4	76.4	68.5	78.4	63.9	71.5	72.7	47.4	51.4	83.9	60.7	63.9	79.9
		4	66.6	81.8	63.0	77.5	64.5	72.6	73.4	46.6	51.4	85.7	60.0	64.5	79.9
		5	66.6	74.5	68.5	76.1	65.2	73.2	73.4	47.0	53.1	87.5	60.7	62.6	80.5
	RANDOM-K (α-tuned)	1	66.7	80.0	68.5	77.0	66.1	72.0	71.2	50.4	52.0	83.9	59.6	63.2	79.6
		2	66.5	76.4	66.7	77.9	65.5	72.6	71.9	47.4	53.7	83.9	60.7	63.0	78.9
		3	66.2	81.8	64.8	75.7	66.5	71.2	71.2	48.1	52.5	85.7	59.3	62.6	79.9
		4	66.5	78.2	68.5	77.0	66.5	72.0	71.9	47.4	53.1	87.5	60.0	62.8	79.3
		5	66.8	76.4	68.5	76.6	66.8	72.3	72.7	48.9	54.2	89.3	60.7	62.3	79.3
	MERIT	–	67.5	81.8	63.0	77.5	67.7	71.5	75.5	45.9	57.6	82.1	62.5	64.5	80.5
Qwen3-VL-4B	BASE	–	64.0	69.1	72.2	76.6	62.6	71.8	74.1	47.8	45.2	75.0	55.8	59.5	77.4
	LORA	–	64.7	76.4	74.1	76.1	64.5	72.0	74.1	46.3	57.1	76.8	54.7	60.8	73.1
	FFT	–	65.0	85.5	75.9	74.8	67.4	69.5	71.9	41.4	70.6	87.5	54.7	58.1	74.0
	FULL-BACKBONE	–	63.5	67.3	70.4	78.8	62.0	71.2	71.9	44.0	45.8	75.0	57.5	57.7	77.7
	ALL-ATTENTION	–	64.1	67.3	72.2	76.6	63.3	72.3	73.4	48.9	44.1	75.0	55.4	59.5	77.4
	RANDOM-K	1	63.1	65.5	72.2	78.8	61.3	71.2	69.8	47.0	44.1	76.8	55.4	57.7	76.5
		2	60.4	60.0	64.8	76.1	59.4	67.5	71.9	41.0	40.1	75.0	53.0	56.4	73.7
		3	60.0	67.3	63.0	75.2	60.7	68.9	71.2	36.9	39.0	76.8	51.6	55.7	73.7
		4	61.2	61.8	72.2	75.2	61.3	67.2	69.1	43.3	39.5	73.2	52.6	58.1	76.2
		5	61.8	61.8	63.0	77.0	60.1	68.9	70.5	40.3	44.6	75.0	54.7	58.1	77.7
	RANDOM-K (α-tuned)	1	59.0	70.9	66.7	71.6	57.2	64.4	65.5	41.8	45.2	67.9	51.2	57.5	69.3
		2	58.5	67.3	63.0	71.6	56.9	63.8	64.7	37.3	44.1	67.9	51.6	59.0	69.3
		3	63.1	69.1	72.2	81.1	63.6	69.2	72.7	42.2	45.8	78.6	53.3	58.4	76.5
		4	63.3	70.9	72.2	78.4	62.6	71.2	73.4	44.4	44.1	76.8	54.0	59.0	75.9
		5	62.8	69.1	72.2	79.3	61.0	70.1	73.4	41.4	42.9	76.8	55.1	59.0	76.2
	MERIT	–	63.6	70.9	72.2	78.8	62.3	71.5	71.2	43.3	46.9	75.0	55.8	58.8	77.1

Table 12: **Full results (Accuracy %) on LongVideoBench.** The columns S2E, S2O, S2A, E2O, O2E, T2E, T2O, and T2A belong to the higher-level category *Perception*, while E3E, O3O, SSS, SOS, SAA, T3E, T3O, TOS, and TAA belong to *Relation*.

Model	Variant	Run	Overall	Perception								Relation								
				S2E	S2O	S2A	E2O	O2E	T2E	T2O	T2A	E3E	O3O	SSS	SOS	SAA	T3E	T3O	TOS	TAA
LongVA-7B	BASE	–	52.5	60.2	48.6	75.0	52.3	58.6	55.4	51.3	54.4	56.4	51.5	35.1	65.4	51.4	43.8	50.0	35.6	43.9
	LORA	–	52.2	60.2	47.2	73.9	52.3	60.9	56.9	52.6	54.4	54.3	48.5	34.0	64.2	50.0	45.2	52.7	31.5	45.1
	FFT	–	40.8	45.2	36.1	67.0	44.6	49.4	50.8	38.2	40.5	45.7	40.9	21.6	38.3	37.5	34.2	43.2	27.4	32.9
	FULL-BACKBONE	–	52.9	62.4	52.8	76.1	50.8	59.8	58.5	51.3	51.9	55.3	48.5	34.0	66.7	54.2	45.2	50.0	35.6	42.7
	ALL-ATTENTION	–	52.6	60.2	48.6	75.0	50.8	58.6	55.4	52.6	55.7	56.4	53.0	33.0	66.7	52.8	43.8	50.0	34.2	43.9
	RANDOM-K	1	52.8	62.4	50.0	75.0	46.2	57.5	53.8	50.0	53.2	59.6	50.0	36.1	65.4	54.2	46.6	50.0	37.0	45.1
		2	52.5	59.1	56.9	75.0	58.5	59.8	55.4	51.3	48.1	50.0	45.5	33.0	67.9	52.8	45.2	54.1	34.2	45.1
		3	52.8	60.2	51.4	75.0	50.8	59.8	60.0	53.9	51.9	55.3	51.5	33.0	65.4	50.0	47.9	51.4	31.5	46.3
		4	52.6	60.2	52.8	76.1	49.2	58.6	60.0	50.0	53.2	53.2	50.0	32.0	65.4	52.8	46.6	50.0	37.0	45.1
		5	53.0	60.2	52.8	76.1	49.2	60.9	55.4	53.9	53.2	55.3	53.0	30.9	65.4	54.2	45.2	54.1	35.6	43.9
	RANDOM-K (α-tuned)	1	52.8	62.4	50.0	75.0	46.2	57.5	53.8	50.0	53.2	59.6	50.0	36.1	65.4	54.2	46.6	50.0	37.0	45.1
		2	52.7	58.1	56.9	76.1	58.5	58.6	56.9	52.6	48.1	51.1	45.5	30.9	65.4	52.8	45.2	55.4	37.0	46.3
		3	52.7	58.1	55.6	75.0	52.3	55.2	60.0	52.6	51.9	54.3	48.5	34.0	66.7	54.2	43.8	51.4	34.2	47.6
		4	52.8	58.1	59.7	73.9	55.4	57.5	60.0	51.3	50.6	54.3	48.5	37.1	63.0	51.4	43.8	52.7	32.9	46.3
		5	53.3	60.2	55.6	75.0	47.7	60.9	56.9	56.6	51.9	55.3	53.0	30.9	65.4	55.6	45.2	55.4	35.6	42.7
	MERIT	–	53.6	62.4	55.6	73.9	50.8	58.6	58.5	55.3	55.7	56.4	48.5	32.0	65.4	51.4	43.8	56.8	38.4	46.3
InternVL3-8B	BASE	–	59.8	68.8	65.3	78.4	73.8	58.6	56.9	57.9	63.3	70.2	57.6	43.3	63.0	54.2	58.9	52.7	45.2	47.6
	LORA	–	58.9	66.7	65.3	77.3	70.8	54.0	55.4	59.2	63.3	66.0	56.1	46.4	61.7	54.2	57.5	51.4	43.8	51.2
	FFT	–	55.7	64.5	55.6	75.0	63.1	51.7	50.8	59.2	54.4	61.7	50.0	43.3	61.7	51.4	61.6	48.6	38.4	52.4
	FULL-BACKBONE	–	59.9	68.8	62.5	78.4	73.8	57.5	60.0	57.9	63.3	71.3	57.6	44.3	64.2	50.0	57.5	55.4	42.5	51.2
	ALL-ATTENTION	–	59.5	64.5	63.9	77.3	72.3	57.5	58.5	60.5	64.6	71.3	59.1	45.4	63.0	54.2	57.5	51.4	43.8	46.3
	RANDOM-K	1	59.5	66.7	65.3	73.9	69.2	56.3	60.0	60.5	62.0	69.1	56.1	45.4	66.7	54.2	58.9	55.4	42.5	48.8
		2	59.8	68.8	62.5	75.0	72.3	57.5	60.0	57.9	62.0	70.2	57.6	42.3	67.9	52.8	60.3	55.4	42.5	51.2
		3	58.7	68.8	62.5	78.4	66.2	58.6	60.0	56.6	60.8	69.1	63.6	42.3	59.3	54.2	53.4	52.7	41.1	48.8
		4	60.1	65.6	65.3	73.9	73.8	56.3	58.5	60.5	64.6	70.2	59.1	45.4	66.7	51.4	60.3	55.4	45.2	48.8
		5	59.3	65.6	62.5	76.1	72.3	56.3	61.5	57.9	62.0	71.3	54.5	44.3	69.1	54.2	56.2	51.4	42.5	48.8
	RANDOM-K (α-tuned)	1	59.8	67.7	63.9	79.5	69.2	57.5	61.5	60.5	64.6	67.0	56.1	42.3	67.9	56.9	57.5	54.1	41.1	48.8
		2	59.9	71.0	66.7	76.1	69.2	57.5	58.5	59.2	65.8	69.1	56.1	43.3	66.7	58.3	58.9	51.4	39.7	48.8
		3	60.0	66.7	63.9	76.1	70.8	57.5	63.1	61.8	62.0	67.0	56.1	43.3	69.1	55.6	58.9	54.1	42.5	51.2
		4	60.1	65.6	63.9	75.0	70.8	58.6	61.5	59.2	64.6	69.1	56.1	43.3	70.4	55.6	57.5	52.7	43.8	52.4
		5	60.4	68.8	65.3	77.3	70.8	56.3	61.5	61.8	64.6	69.1	56.1	44.3	70.4	56.9	57.5	52.7	43.8	48.8
	MERIT	–	60.3	69.9	61.1	75.0	69.2	59.8	63.1	61.8	67.1	68.1	54.5	43.3	65.4	54.2	61.6	54.1	39.7	54.9
Qwen3-VL-4B	BASE	–	61.9	66.7	62.5	81.8	63.1	70.1	64.6	61.8	74.7	60.6	63.6	42.3	72.8	63.9	56.2	51.4	41.1	54.9
	LORA	–	61.1	66.7	63.9	79.5	64.6	69.0	64.6	57.9	70.9	58.5	62.1	48.5	67.9	61.1	57.5	56.8	39.7	48.8
	FFT	–	55.3	63.4	50.0	73.9	61.5	55.2	50.8	52.6	65.8	56.4	56.1	47.4	58.0	50.0	56.2	52.7	37.0	50.0
	FULL-BACKBONE	–	60.4	63.4	61.1	76.1	66.2	65.5	60.0	57.9	70.9	68.1	65.2	46.4	70.4	55.6	57.5	50.0	37.0	52.4
	ALL-ATTENTION	–	62.0	64.5	65.3	80.7	63.1	69.0	64.6	63.2	74.7	61.7	63.6	42.3	74.1	62.5	57.5	54.1	41.1	52.4
	RANDOM-K	1	60.2	66.7	61.1	78.4	63.1	66.7	64.6	60.5	70.9	66.0	65.2	42.3	71.6	54.2	56.2	50.0	34.2	50.0
		2	59.8	66.7	54.2	77.3	70.8	64.4	60.0	59.2	68.4	62.8	60.6	42.3	72.8	58.3	57.5	48.6	39.7	52.4
		3	59.2	62.4	62.5	77.3	67.7	58.6	60.0	57.9	65.8	64.9	63.6	38.1	70.4	59.7	57.5	51.4	38.4	52.4
		4	60.0	62.4	59.7	76.1	60.0	63.2	61.5	59.2	65.8	64.9	63.6	46.4	72.8	59.7	57.5	51.4	41.1	52.4
		5	59.8	64.5	61.1	75.0	67.7	63.2	61.5	59.2	68.4	64.9	57.6	39.2	75.3	61.1	54.8	48.6	41.1	53.7
	RANDOM-K (α-tuned)	1	60.7	66.7	65.3	78.4	61.5	62.1	60.0	61.8	67.1	67.0	68.2	42.3	76.5	58.3	54.8	54.1	35.6	51.2
		2	59.9	64.5	63.9	75.0	64.6	63.2	58.5	57.9	68.4	67.0	59.1	44.3	72.8	59.7	52.1	55.4	37.0	52.4
		3	60.7	64.5	62.5	78.4	61.5	66.7	61.5	60.5	68.4	62.8	59.1	42.3	77.8	63.9	54.8	52.7	41.1	52.4
		4	60.7	65.6	65.3	77.3	63.1	62.1	61.5	63.2	64.6	63.8	63.6	44.3	75.3	59.7	54.8	54.1	41.1	52.4
		5	60.8	66.7	65.3	79.5	61.5	62.1	61.5	59.2	69.6	67.0	63.6	40.2	75.3	61.1	56.2	52.7	37.0	53.7
	MERIT	–	60.6	64.5	68.1	78.4	64.6	62.1	56.9	53.9	67.1	61.7	62.1	46.4	74.1	62.5	56.2	58.1	37.0	54.9

Table 13: **Full results (Accuracy %) on LVBench.** Overall: Overall performance; ER: Entity Recognition; EU: Event Understanding; KIR: Key Information Retrieval; TG: Temporal Grounding; Sum: Summarization; Rea: Reasoning.

Model	Variant	Run	Overall	ER	EU	KIR	TG	Sum	Rea
LongVA-7B	BASE	–	37.6	38.3	35.1	37.1	29.5	39.7	45.3
	LORA	–	37.7	39.3	35.4	35.4	30.9	34.5	45.3
	FFT	–	35.4	36.5	32.8	35.7	31.4	24.1	41.8
	FULL-BACKBONE	–	38.1	38.4	36.2	37.8	32.3	37.9	44.8
	ALL-ATTENTION	–	38.0	39.0	35.5	36.4	31.4	41.4	46.3
	RANDOM-K	1	38.6	39.4	36.2	37.5	29.5	37.9	46.8
		2	37.1	37.5	35.9	35.7	30.5	39.7	42.3
		3	37.7	39.7	36.0	37.1	30.9	34.5	44.3
		4	37.4	38.7	35.1	36.8	32.3	41.4	44.8
		5	38.0	39.0	35.2	36.8	30.9	41.4	43.8
	RANDOM-K (α-tuned)	1	38.6	39.4	36.2	37.5	29.5	37.9	46.8
		2	36.7	37.8	34.8	34.7	30.5	43.1	42.3
		3	37.1	38.4	35.5	36.8	31.8	39.7	44.8
		4	38.0	39.1	37.1	35.4	31.8	39.7	46.3
		5	38.3	39.1	35.4	38.1	32.3	41.4	46.3
	MERIT	–	39.7	41.1	37.4	37.5	33.2	43.1	46.8
InternVL3-8B	BASE	–	44.8	45.1	43.0	47.8	37.3	29.3	51.2
	LORA	–	44.7	44.3	44.5	47.8	35.0	27.6	52.2
	FFT	–	44.4	44.6	43.9	45.4	37.3	22.4	50.2
	FULL-BACKBONE	–	44.7	44.3	45.0	47.1	34.1	32.8	50.7
	ALL-ATTENTION	–	44.7	44.6	43.3	48.1	36.4	31.0	50.2
	RANDOM-K	1	44.3	44.2	43.7	46.4	36.4	31.0	48.3
		2	43.8	43.0	43.9	45.4	35.9	32.8	50.2
		3	44.5	43.4	44.0	47.8	35.9	36.2	50.2
		4	45.0	45.3	43.9	47.8	36.8	31.0	51.7
		5	43.2	42.8	42.3	47.1	35.5	27.6	48.8
	RANDOM-K (α-tuned)	1	44.8	44.6	44.2	48.1	38.6	32.8	48.8
		2	44.2	44.5	43.6	46.7	37.7	27.6	48.3
		3	44.0	45.1	42.8	47.1	37.3	25.9	48.3
		4	44.6	44.9	43.6	48.1	35.5	31.0	48.3
		5	44.1	44.6	42.8	47.4	35.5	29.3	46.8
	MERIT	–	45.0	44.8	44.2	47.4	35.5	29.3	53.2
Qwen3-VL-4B	BASE	–	39.4	39.6	39.6	39.5	30.9	29.3	39.3
	LORA	–	40.4	39.9	39.7	39.9	36.4	29.3	42.3
	FFT	–	40.3	42.5	39.4	41.2	30.9	20.7	43.3
	FULL-BACKBONE	–	34.9	35.2	34.9	32.6	30.0	27.6	33.3
	ALL-ATTENTION	–	43.7	43.6	43.9	45.4	35.9	32.8	43.3
	RANDOM-K	1	41.6	43.0	39.6	43.3	34.1	24.1	44.3
		2	38.3	39.4	35.9	37.8	34.1	19.0	40.8
		3	38.3	39.7	36.5	41.9	34.5	20.7	38.8
		4	39.1	38.1	38.2	40.9	38.2	34.5	41.8
		5	37.9	39.3	36.9	37.5	30.9	22.4	36.8
	RANDOM-K (α-tuned)	1	38.3	39.3	37.1	42.3	31.4	20.7	40.8
		2	40.9	41.8	39.1	43.6	32.7	27.6	42.8
		3	40.2	40.0	40.2	43.0	33.2	24.1	40.3
		4	40.7	42.4	38.8	43.0	33.6	20.7	43.3
		5	40.0	41.1	37.9	42.3	31.4	22.4	40.3
	MERIT	–	40.6	40.6	39.9	41.2	32.3	25.9	46.3

Table 14: **Full results on MMBench-Video.** CP: Coarse Perception; FP-S: Fine-grained Perception (Single-Instance); FP-C: Fine-grained Perception (Cross-Instance); HL: Hallucination; LR: Logic Reasoning; AR: Attribute Reasoning; RR: Relation Reasoning; CSR: Common Sense Reasoning; TR: Temporal Reasoning. All scores are based on a 3-grade marking scheme: 0 for worst, 3 for best.

Model	Variant	Run	Overall	CP	FP-S	FP-C	HL	LR	AR	RR	CSR	TR
LongVA-7B	BASE	–	1.27	1.55	1.24	1.06	1.42	0.66	1.61	1.30	1.22	1.05
	LORA	–	1.24	1.58	1.18	1.11	1.50	0.61	1.59	1.22	1.21	1.04
	FFT	–	1.15	1.43	1.10	0.99	1.45	0.58	1.35	1.14	1.01	0.98
	FULL-BACKBONE	–	1.25	1.56	1.23	1.13	1.56	0.63	1.42	1.32	1.20	1.08
	ALL-ATTENTION	–	1.28	1.56	1.27	1.14	1.26	0.60	1.65	1.30	1.20	1.08
	RANDOM-K	1	1.28	1.60	1.22	1.13	1.60	0.61	1.58	1.27	1.10	1.06
		2	1.24	1.56	1.18	1.09	1.47	0.61	1.45	1.27	1.17	1.05
		3	1.25	1.56	1.22	1.14	1.52	0.58	1.57	1.36	1.15	1.08
		4	1.25	1.61	1.24	1.13	1.56	0.59	1.57	1.24	1.19	1.08
		5	1.23	1.51	1.20	1.11	1.58	0.68	1.48	1.30	1.23	1.09
	RANDOM-K (α-tuned)	1	1.28	1.60	1.22	1.13	1.60	0.61	1.58	1.27	1.10	1.06
		2	1.25	1.59	1.18	1.12	1.58	0.55	1.51	1.29	1.14	1.07
		3	1.27	1.53	1.20	1.13	1.56	0.54	1.51	1.33	1.21	1.09
		4	1.23	1.59	1.24	1.15	1.56	0.63	1.61	1.32	1.17	1.07
		5	1.24	1.58	1.22	1.08	1.58	0.60	1.58	1.28	1.14	1.07
	MERIT	–	1.29	1.61	1.25	1.15	1.40	0.52	1.57	1.30	1.16	1.09
InternVL3-8B	BASE	–	1.68	1.79	1.84	1.54	0.32	1.29	1.78	1.77	1.52	1.44
	LORA	–	1.71	1.77	1.87	1.50	0.35	1.33	1.86	1.88	1.51	1.42
	FFT	–	1.69	1.82	1.88	1.51	0.37	1.41	1.86	1.87	1.65	1.41
	FULL-BACKBONE	–	1.71	1.74	1.85	1.46	0.37	1.32	1.85	1.81	1.54	1.44
	ALL-ATTENTION	–	1.71	1.81	1.88	1.54	0.32	1.32	1.78	1.86	1.58	1.44
	RANDOM-K	1	1.71	1.89	1.87	1.49	0.37	1.36	1.85	1.83	1.60	1.44
		2	1.71	1.81	1.86	1.53	0.40	1.26	1.86	1.89	1.68	1.44
		3	1.69	1.77	1.86	1.56	0.32	1.24	1.86	1.75	1.59	1.40
		4	1.71	1.81	1.90	1.51	0.39	1.35	1.89	1.88	1.69	1.44
		5	1.70	1.85	1.89	1.49	0.34	1.33	1.79	1.83	1.53	1.45
	RANDOM-K (α-tuned)	1	1.71	1.77	1.89	1.53	0.32	1.33	1.87	1.79	1.70	1.46
		2	1.70	1.85	1.89	1.56	0.32	1.31	1.85	1.88	1.54	1.43
		3	1.71	1.84	1.89	1.52	0.32	1.22	1.85	1.80	1.60	1.45
		4	1.70	1.79	1.87	1.56	0.37	1.37	1.83	1.82	1.59	1.42
		5	1.69	1.85	1.86	1.52	0.37	1.31	1.88	1.85	1.65	1.45
	MERIT	–	1.72	1.86	1.86	1.61	0.37	1.27	1.85	1.81	1.73	1.46
Qwen3-VL-4B	BASE	–	1.50	1.76	1.59	1.22	0.66	1.47	1.76	1.42	1.48	1.08
	LORA	–	1.56	1.87	1.63	1.28	0.92	1.50	1.95	1.79	1.41	1.24
	FFT	–	1.53	1.83	1.65	1.12	0.74	1.32	1.73	1.61	1.33	1.11
	FULL-BACKBONE	–	1.56	1.91	1.59	1.25	0.87	1.23	1.88	1.49	1.65	1.26
	ALL-ATTENTION	–	1.59	1.84	1.66	1.27	0.85	1.57	1.82	1.48	1.58	1.27
	RANDOM-K	1	1.59	1.91	1.66	1.23	0.94	1.55	1.86	1.44	1.64	1.28
		2	1.56	1.85	1.59	1.29	0.87	1.33	1.81	1.66	1.43	1.30
		3	1.41	1.80	1.46	1.08	0.68	1.26	1.59	1.36	1.35	1.08
		4	1.59	1.84	1.67	1.27	0.76	1.42	1.80	1.65	1.78	1.36
		5	1.62	1.91	1.67	1.28	1.00	1.35	1.80	1.58	1.64	1.34
	RANDOM-K (α-tuned)	1	1.55	1.84	1.64	1.29	1.15	1.38	1.63	1.56	1.75	1.22
		2	1.57	1.75	1.58	1.25	1.11	1.34	1.77	1.46	1.38	1.35
		3	1.72	1.84	1.58	1.17	0.89	1.60	1.94	1.65	1.52	1.34
		4	1.49	1.88	1.59	1.35	1.05	1.32	1.99	1.32	1.70	1.16
		5	1.44	1.83	1.54	1.27	0.82	1.39	1.90	1.68	1.30	1.24
	MERIT	–	1.66	1.94	1.72	1.29	0.97	1.63	1.88	1.58	1.65	1.38

Table 15: **Full results (Accuracy %) on Video-Holmes.** CTI: Core Theme Inference; IMC: Intention & Motive Chaining; MHR: Multimodal Hint Reasoning; PAR: Physical Anomaly Reasoning; SR: Social Reasoning; TA: Timeline Analysis; TCI: Temporal Causal Inference.

Model	Variant	Run	Overall	CTI	IMC	MHR	PAR	SR	TA	TCI
LongVA-7B	BASE	–	15.2	9.3	18.8	18.1	11.3	24.3	3.0	15.8
	LORA	–	14.5	9.6	17.0	19.0	10.3	24.7	1.5	13.2
	FFT	–	4.7	2.6	8.0	5.1	2.6	7.9	0.5	4.4
	FULL-BACKBONE	–	14.6	13.0	18.1	16.0	13.9	22.9	1.0	12.8
	ALL-ATTENTION	–	15.4	12.2	17.4	18.7	15.5	23.3	2.0	13.9
	RANDOM-K	1	16.7	12.2	22.5	17.5	13.4	26.7	1.0	17.6
		2	10.0	7.0	12.3	11.4	7.7	18.2	1.5	8.1
		3	13.7	13.0	13.4	14.8	11.3	26.4	2.5	9.5
		4	19.8	16.3	25.4	24.4	12.9	20.5	12.0	21.6
		5	12.3	8.9	13.4	13.0	12.4	19.5	5.5	11.0
	RANDOM-K (α-tuned)	1	16.7	12.2	22.5	17.5	13.4	26.7	1.0	17.6
		2	13.7	10.7	14.5	12.0	14.4	27.7	2.5	10.6
		3	15.2	13.7	17.4	15.7	12.9	26.0	2.5	13.6
		4	19.8	17.0	22.5	19.9	17.0	33.6	3.5	18.7
		5	13.7	13.0	13.8	15.4	12.4	22.9	4.5	10.3
	MERIT	–	18.9	17.0	23.6	20.5	17.0	29.8	2.0	16.1
InternVL3-8B	BASE	–	36.5	34.8	42.4	34.0	35.6	44.5	31.0	31.5
	LORA	–	36.2	36.3	36.2	31.3	34.0	42.5	37.5	35.9
	FFT	–	35.7	34.4	37.3	32.5	32.0	44.5	36.0	32.2
	FULL-BACKBONE	–	36.3	33.7	39.5	35.8	29.4	48.6	28.5	33.7
	ALL-ATTENTION	–	36.2	33.0	39.5	33.4	35.1	45.2	31.5	34.1
	RANDOM-K	1	37.4	35.9	40.6	36.1	32.5	51.0	26.5	34.1
		2	36.1	36.7	39.9	32.2	33.5	48.3	29.0	30.4
		3	35.1	32.2	36.2	34.0	32.0	47.3	31.5	30.0
		4	37.4	37.0	38.8	34.0	33.0	50.3	32.0	33.7
		5	37.4	34.4	40.6	34.6	36.6	51.0	33.5	29.3
	RANDOM-K (α-tuned)	1	35.9	34.4	36.6	37.3	31.4	49.3	28.0	29.7
		2	37.1	35.9	43.1	34.0	31.4	52.4	25.5	31.9
		3	36.1	36.3	38.4	32.8	32.5	49.0	23.5	35.9
		4	37.0	36.3	38.0	34.6	32.0	49.7	31.0	33.7
		5	36.4	36.3	38.4	34.0	33.0	51.0	26.5	31.5
	MERIT	–	37.2	34.1	39.9	37.0	37.1	49.7	30.5	29.3
Qwen3-VL-4B	BASE	–	31.5	33.7	35.1	31.0	19.6	38.7	35.5	24.2
	LORA	–	32.0	29.3	35.1	28.9	25.8	38.0	43.5	24.5
	FFT	–	30.7	28.1	40.2	28.3	24.7	33.6	39.0	21.6
	FULL-BACKBONE	–	31.2	33.0	36.2	29.5	25.8	41.8	27.5	21.6
	ALL-ATTENTION	–	31.9	29.6	34.1	29.8	21.1	45.9	37.0	23.4
	RANDOM-K	1	30.3	28.5	33.3	31.3	21.6	44.9	31.0	17.9
		2	31.3	33.0	36.2	29.5	25.8	42.5	27.5	21.6
		3	30.2	27.0	34.1	27.7	21.6	43.8	36.0	19.8
		4	32.0	31.1	34.8	29.8	27.8	46.9	29.0	21.6
		5	31.3	30.4	35.9	30.7	26.3	40.8	36.5	17.9
	RANDOM-K (α-tuned)	1	32.1	29.6	33.7	30.4	24.2	45.5	35.5	23.8
		2	30.8	28.9	32.2	28.6	22.2	43.8	34.0	23.8
		3	32.6	31.9	35.5	30.4	23.7	46.2	37.0	21.6
		4	31.7	30.0	34.4	29.8	24.7	43.8	34.5	22.7
		5	31.2	29.6	35.1	30.1	25.3	41.1	36.0	20.5
	MERIT	–	32.9	32.6	38.4	31.9	23.2	45.2	32.0	23.1