EDITING WHAT THE VIDEO CONTRIBUTES: VISUAL INTERVENTION MEETS FISHER GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Successful editing in Video-LLM can mask *Visual Blindness*: the correction may rely on text-side shortcuts rather than videos. We study this ambiguity through a structure-preserving Visual-Null intervention, revealing that substantial editing gains persist without informative video. We translate this behavioral contrast into Fisher geometry to isolate editing directions beyond the blind response. The resulting Fisher-orthogonal residual defines *video-grounded editability*, capturing both separation from the Visual-Null response and editing target controllability. This formulation yields VIGOR (Visual-Intervention Guided Orthogonal Residualization), a closed-form, minimum-information update that improves the video-conditioned objective while preserving the Visual-Null objective to first order. We prove that greater video-grounded editability lowers the minimum functional cost of a prescribed target improvement, informing principled layer selection within the locate-then-edit paradigm. Across four Video-LLMs on VMEB, VIGOR achieves the highest average editing score among evaluated methods, combining near-perfect reliability with near-zero Visual-Null gains. It also largely preserves general video understanding and representation structure after 100 sequential edits. *Our code is provided in the supplementary materials; all data will be released soon on Github.*

1 INTRODUCTION

Large language models (LLMs) have expanded from language processing to multimodal and video understanding (Touvron et al., 2023; Zhao et al., 2023; Liu et al., 2023; Lin et al., 2024; Tang et al., 2025). However, the knowledge they acquire is encoded in model parameters and remains fixed after training, making their predictions misaligned with an evolving world (Jiang et al., 2026). Model editing offers an efficient alternative to retraining, aiming to update specific knowledge or behaviors while preserving unrelated capabilities (Meng et al., 2022; Fang et al., 2025; Chen et al., 2025b).

Recent work extends model editing to MLLMs and Video-LLMs, where target responses depend on both textual queries and visual evidence (Cheng et al., 2023; Li et al., 2026; Zhang et al., 2026). However, improving the likelihood of a target response does not necessarily establish its dependence on video (Zhang et al., 2026). Given an editing request $(x^v, x^q, y^\star)$, a standard objective seeks an update ΔW that increases $p_{\theta_{\Delta W}}(y^\star \mid x^v, x^q)$. This objective does not distinguish a video-dependent correction from a text-side shortcut that reinforces the question–answer association. Indeed, prior work reports that mainstream editors can retain high editing success after informative visual content is removed, a failure mode termed *Visual Blindness* (Zhang et al., 2026). This raises a central question: *how can we direct an edit toward what the video contributes beyond the visual blindness?*

To bridge this gap, we first study the anatomy of Video-LLM editing (Section 3.1). We introduce a structure-preserving *Visual-Null* intervention that removes instance-specific visual content while retaining the textual query and multimodal input structure. Comparing the editing gains under the original input (x^v, x^q) and its Visual-Null counterpart $(x^{v,\varnothing}, x^q)$ quantifies how much of an edit remains effective without informative visual evidence. We refer to the local editing response induced under Visual-Null as the *blind response*. Our analysis shows that conventional editors can still achieve substantial gains under Visual-Null, often comparable to those obtained with the original video.

Visual-Null exposes this ambiguity at the behavioral level, but does not specify how to construct an update that isolates the response component beyond the blind pathway. As illustrated in Figure 1, we

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.

therefore lift the Visual–Null contrast into the local geometry of parameter updates (Section 3.2). Considering Euclidean distance in parameter space does not reliably reflect functional change (Martens, 2020), we introduce the *Fisher geometry* induced by the model's predictive distribution (Xu et al., 2026). Under this geometry, the Visual-Null response defines a *blind tangent direction*, while the Fisher-orthogonal residual isolates the component of the video-conditioned response beyond this tangent. We quantify the strength of this residual as *video-grounded editability*, jointly capturing separation from the Visual-Null response and target controllability. Its depth-wise profile provides a principle for edit localization.

Building on this analysis, we propose **VIGOR** (**V**isual-**I**ntervention **G**uided **O**rthogonal **R**esidualization), which turns the Visual–Null contrast into a geometric constraint on parameter updates. VIGOR removes the blind component through Fisher-orthogonal residualization, then derives a closed-form, minimum-information update that achieves the desired local improvement with minimal disturbance to the model's distribution. We further prove that, for a fixed target improvement, greater video-grounded editability implies a lower minimum functional cost, providing a principled basis for the locate-then-edit paradigm. Our main contributions are:

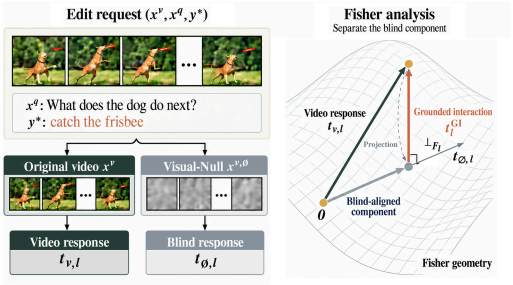


Figure 1: **Editing what the video contributes.** Visual-Null reveals the blind response, while Fisher-orthogonal decomposition isolates the residual beyond this response from the full response.

- We introduce a structure-preserving *Visual-Null* intervention to diagnose blind editing gains in Video-LLM editing, revealing that substantial gains achieved by conventional editors remain accessible without informative visual evidence (Section 3.1).
- We develop a Fisher-geometric characterization of *video-grounded editability*, which isolates the functionally effective editing capacity beyond the blind response and reveals where separation from the Visual-Null response and target controllability coexist across model depth (Section 3.2).
- We propose **VIGOR**, which removes the Visual-Null editing response through Fisher-orthogonal residualization and performs a minimum-information parameter update. We prove that greater video-grounded editability implies a lower minimum functional cost for a fixed target improvement, providing a theoretical basis for layer localization (Section 4).
- Extensive experiments across four Video-LLMs validate VIGOR as a principled framework for suppressing blind gains in Video-LLM editing, demonstrating effective correction, mitigating Visual Blindness, and preserving general capabilities under sequential edits (Section 5).

2 Related Work

Video Large Language Models. Video-LLMs extend LLMs to video by integrating visual representations with language reasoning (Li et al., 2025). Early approaches connected pretrained visual encoders to frozen language backbones, laying the foundation for video–language understanding (Alayrac et al., 2022; Yang et al., 2022). Instruction-tuned models, such as VideoChat (Li et al., 2025), Video-LLaMA (Zhang et al., 2023), and Video-LLaVA (Lin et al., 2024), subsequently improved video–language alignment. More recent models, including LLaVA-NeXT-Video (Zhang et al., 2024) and the Qwen-VL family (Bai et al., 2025a), further advance video understanding through stronger visual representations and temporal modeling.

Multimodal LLM Editing. Model editing revises specific knowledge or behaviors through parameter updates, external memory, or in-context learning (Meng et al., 2022; Mitchell et al., 2022b; Zheng et al., 2023). Early MLLM studies adapt text-based editors, whose effectiveness and limitations are systematically evaluated by MMEdit (Cheng et al., 2023). Subsequent designs include attribution-guided visual unit localization and modification in VisEdit (Chen et al., 2025b), gated dual-branch editing in DualEdit (Shi et al., 2025), and visual–textual expert composition for lifelong editing in LiveEdit (Chen et al., 2025a), while M-ORE analyzes cross-modal contamination (Li et al., 2026). VLKEB further broadens multimodal editing evaluation (Huang et al., 2024). These efforts primarily address static images; VMEB extends editing to video and reveals *Visual Blindness*, exposing the gap between behavioral correction and visual dependence (Zhang et al., 2026).

3 THE ANATOMY OF VIDEO-LLM EDITING

Video Large Language Model. We consider a Video-LLM f_θ with a vision encoder $\mathcal{E}_v$, projector $\mathcal{P}$, and LLM $\mathcal{M}$. Given (x^v, x^q), where x^v is the video input and x^q is the text query, the model autoregressively predicts $p(y \mid x^v, x^q)$ from embeddings $\boldsymbol{H} = [\mathcal{P}(\mathcal{E}_v(x^v)), \mathrm{Embed}(x^q)]$. This process can be viewed as learning a mapping from the joint vision-language representation space to the text answer token distribution space. At transformer layer l in the LLM, the FFN computation can be written as (Meng et al., 2022; Fang et al., 2025):

$$\boldsymbol{h}_l = \boldsymbol{h}_{l-1} + \boldsymbol{a}_l + \boldsymbol{m}_l, \qquad \boldsymbol{m}_l = \boldsymbol{W}_l^{\text{out}}\, \sigma\big(\boldsymbol{W}_l^{\text{in}}\, \gamma(\boldsymbol{h}_{l-1} + \boldsymbol{a}_l)\big)\,. \tag{1}$$

where $\boldsymbol{a}_l$ and $\boldsymbol{m}_l$ denote the attention and FFN outputs, $\gamma(\cdot)$ is LayerNorm, and $\sigma(\cdot)$ is a nonlinearity. Under the FFN key–value view (Geva et al., 2021), intermediate activations form a memory pair:

$$\boldsymbol{k}_l \triangleq \sigma\big(\boldsymbol{W}_l^{\text{in}}\, \gamma(\boldsymbol{h}_{l-1} + \boldsymbol{a}_l)\big)\,, \qquad \boldsymbol{v}_l \triangleq \boldsymbol{m}_l = \boldsymbol{W}_l^{\text{out}} \boldsymbol{k}_l. \tag{2}$$

Here, $\boldsymbol{k}_l$ serves as the input-dependent key, and $\boldsymbol{v}_l$ represents the corresponding value. In this view, the FFN implements an associative memory that retrieves $\boldsymbol{v}_l$ based on $\boldsymbol{k}_l$ and writes it to the residual stream. Many parameter-modifying editors rewrite knowledge by updating $\boldsymbol{W}_l^{\text{out}}$. For simplicity, we use $\boldsymbol{W}_l$ to refer to $\boldsymbol{W}_l^{\text{out}}$ in the following.

Model Editing. Model editing aims to modify specific parameters $\boldsymbol{W}$ of a Video-LLM to alter its output distribution toward a desired response, while preserving the original behavior encoded in the remaining parameters $\theta \setminus \{\boldsymbol{W}\}$. Specifically, a video knowledge editing request can be defined as $(x^v, x^q, y^\star)$, where $y^\star$ denotes the desired target response of the model. Based on this formulation, the editing objective can be expressed as:

$$\Delta \boldsymbol{W} = \arg\min_{\Delta \boldsymbol{W}} \mathcal{L}_v(\theta_{\Delta \boldsymbol{W}}; x^v, x^q) = \arg\min_{\Delta \boldsymbol{W}} -\frac{1}{T} \sum_{t=1}^{T} \log p_{\theta_{\Delta \boldsymbol{W}}}\left(y_t^\star \mid x^v, x^q, y_{<t}^\star\right). \tag{3}$$

where $\theta_{\Delta \boldsymbol{W}}$ denotes the edited model with parameter update $\Delta \boldsymbol{W}$. A successful knowledge editor should satisfy three requirements: the edited model should produce the target response for the editing request (*reliability*), generalize the edit to semantically equivalent inputs (*generality*), and preserve its original behavior on unrelated inputs (*locality*). More detailed editing procedures, discussions, and evaluation metric definitions are provided in Appendix A.2.

These concepts were originally developed for text-only LLMs. When extended to heterogeneous multimodal scenarios, they suffer from *Visual Blindness*. In the following, we analyze this phenomenon in detail and attribute the contribution of video inputs to reveal their role in the model editing.

3.1 EDITING ATTRIBUTION VIA VISUAL INTERVENTION

Under the key–value view, minimizing $\mathcal{L}_v$ in Eq.(3) does not uniquely determine how the target association is retrieved. An editor may exhibit two qualitatively different editing behaviors:

- **Video-grounded Editing.**[1] The editor updates the association from (x^v, x^q) to $y^\star$ through a video-conditioned response component rather than through a shortcut accessible from x^q alone.
- **Text-side Shortcut.** The editor strengthens the association from x^q to $y^\star$, enabling target retrieval without relying on x^v or the video–question interaction. Consequently, the edit-induced improvement can remain accessible even under Visual-Null.

Both mechanisms can achieve comparable performance under standard editing metrics, which fail to distinguish visually grounded edits from *Visual Blindness*. To disentangle these two behaviors, we introduce a structure-preserving Visual-Null intervention: $x^{v,\varnothing} = \mathrm{Null}(x^v)$, which removes instance-specific visual content while preserving the original multimodal structure and textual query. Under this counterfactual input, we define the corresponding Visual-Null objective as

$$\mathcal{L}_\varnothing(\theta; x^{v,\varnothing}, x^q) = -\frac{1}{T} \sum_{t=1}^{T} \log p_\theta\left(y_t^\star \mid x^{v,\varnothing}, x^q, y_{<t}^\star\right). \tag{4}$$

[1]Throughout this work, an edit is considered more *video-grounded* when less of its induced gain remains under Visual-Null, i.e., when the editing effect is more specific to the original video-conditioned response and less attributable to the blind pathway exposed by Visual-Null. This operational definition does not imply semantic causality under arbitrary video counterfactuals.

Next, we measure the loss reduction under the original video input and the Visual-Null input:

$$\Delta_{\text{full}} = \mathcal{L}_v(\theta; x^v, x^q) - \mathcal{L}_v(\theta_{\Delta \boldsymbol{W}}; x^v, x^q),$$
$$\Delta_{\text{null}} = \mathcal{L}_\varnothing(\theta; x^{v,\varnothing}, x^q) - \mathcal{L}_\varnothing(\theta_{\Delta \boldsymbol{W}}; x^{v,\varnothing}, x^q). \tag{5}$$

Here, Δ_{full} measures the edit gain with the original video, whereas Δ_{null} measures the gain retained under Visual-Null. Their contrast operationalizes video-grounding: a video-dependent edit should yield $\Delta_{\text{full}} > 0$ but $\Delta_{\text{null}} \approx 0$. In contrast, $\Delta_{\text{null}} \approx \Delta_{\text{full}}$ indicates that the edit remains effective without instance-specific visual evidence, revealing a blind editing response.

As shown in Figure 2, we apply the Visual-Null intervention to FT (Lang.) as a representative conventional editor on Qwen3-VL and Qwen2.5-VL-7B. Despite removing informative visual evidence, $\Delta_{\text{null}} \geq \Delta_{\text{full}}$ for 72% and 69% of edits, respectively, with aggregate Visual-Null gains reaching $1.40\times$ the original-video gains on both backbones. This persistence exposes *Visual Blindness*: conventional editing objectives can be satisfied without establishing video-dependent target associations. Experimental details and more complementary cross-architecture evidence are provided in Appendix B.1. Although Visual-Null exposes this behavioral discrepancy, it does not reveal how $\Delta \boldsymbol{W}$ induces it. We therefore lift the contrast to local parameter space to disentangle blind and video-grounded editing.

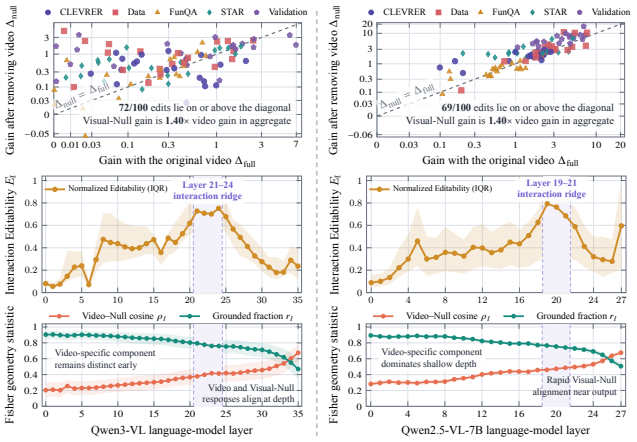


Figure 2: **Diagnosing Visual Blindness and understanding video-grounded editability in Video-LLM editing.**

3.2 The Geometry of Video-Grounded Editability

We consider a candidate FFN layer l with editable parameters $\boldsymbol{W}_l$. The local editing responses induced by the video-conditioned and Visual-Null objectives can be characterized by their gradients:

$$\boldsymbol{g}_{v,l} = \nabla_{\boldsymbol{W}_l} \mathcal{L}_v(\theta; x^v, x^q), \qquad \boldsymbol{g}_{\varnothing,l} = \nabla_{\boldsymbol{W}_l} \mathcal{L}_\varnothing(\theta; x^{v,\varnothing}, x^q). \tag{6}$$

The overlap between these gradients reveals how strongly the video-conditioned editing response aligns with the blind response. To assess this overlap, we require a geometry that accounts for the model's functional sensitivity: parameter perturbations of equal Euclidean magnitude can induce markedly different changes in its predictive distribution.[2] We therefore adopt Fisher geometry, which measures update cost through the local change in model predictions:

$$\mathbb{E}_{x \sim \mathcal{D}_{\text{ref}}} D_{\text{KL}}\left(p_\theta(\cdot \mid x) \| p_{\theta_{\Delta \boldsymbol{W}_l}}(\cdot \mid x)\right) = \frac{1}{2}(\Delta \boldsymbol{W}_l)^\top \boldsymbol{F}_l(\Delta \boldsymbol{W}_l) + o\big(\|\Delta \boldsymbol{W}_l\|_2^2\big)\,. \tag{7}$$

Here, $\boldsymbol{F}_l$ is the Fisher information matrix for $\boldsymbol{W}_l$ under the reference distribution $\mathcal{D}_{\text{ref}}$. In practice, we use a damped empirical approximation, with $\boldsymbol{F}_l \succ 0$. We map each gradient to its Fisher-scaled response direction, $\boldsymbol{t}_{a,l} = \boldsymbol{F}_l^{-1} \boldsymbol{g}_{a,l}$, and quantify functional overlap through the Fisher inner product:

$$\kappa_{a,b}^{(l)} = \langle \boldsymbol{t}_{a,l}, \boldsymbol{t}_{b,l} \rangle_{\boldsymbol{F}_l} = \boldsymbol{g}_{a,l}^\top \boldsymbol{F}_l^{-1} \boldsymbol{g}_{b,l}, \qquad a, b \in \{v, \varnothing\}. \tag{8}$$

We restrict attention to layers with $\kappa_{v,v}^{(l)} > 0$ and $\kappa_{\varnothing,\varnothing}^{(l)} > 0$, so that subsequent projections and normalizations are well-defined. Under this view, the Visual-Null direction $\boldsymbol{t}_{\varnothing,l}$ defines a *blind tangent*, along whose negative direction the target loss decreases without informative visual evidence. To remove this component from the video-conditioned response, we subtract the Fisher-orthogonal projection of $\boldsymbol{t}_{v,l}$ onto $\boldsymbol{t}_{\varnothing,l}$:

$$\boldsymbol{t}_l^{\text{GI}} = \boldsymbol{F}_l^{-1}\left(\boldsymbol{g}_{v,l} - \frac{\kappa_{v,\varnothing}^{(l)}}{\kappa_{\varnothing,\varnothing}^{(l)}} \boldsymbol{g}_{\varnothing,l}\right), \qquad \boldsymbol{g}_{\varnothing,l}^\top \boldsymbol{t}_l^{\text{GI}} = 0. \tag{9}$$

[2] Euclidean geometry measures how much parameters change, but not how strongly those changes affect model predictions. Fisher geometry accounts for this distinction by assigning greater cost to directions that induce larger local changes in the predictive distribution.

We refer to $\boldsymbol{t}_l^{\mathrm{GI}}$ as the *grounded interaction tangent*. By construction, it is functionally orthogonal to the Visual-Null response, so the Visual-Null objective exhibits no first-order change along this tangent. Therefore, it isolates the component of the video-conditioned editing response that cannot be explained by the blind pathway. Next, to characterize how this residual response evolves across Video-LLM layer depth, we define:

$$\rho_l \triangleq \frac{\kappa_{v,\varnothing}^{(l)}}{\sqrt{\kappa_{v,v}^{(l)}\kappa_{\varnothing,\varnothing}^{(l)}}}, \quad r_l \triangleq 1-\rho_l^2, \quad E_l \triangleq \left\|\boldsymbol{t}_l^{\mathrm{GI}}\right\|_{\boldsymbol{F}_l}^2 = \kappa_{v,v}^{(l)} r_l. \tag{10}$$

Here, ρ_l measures Fisher alignment between the video-conditioned and Visual-Null responses, and r_l is the response-energy fraction retained after removing the blind component. A large r_l indicates relative visual specificity, but not necessarily sufficient strength to alter the target response. We therefore define *Grounded Interaction Editability* E_l, combining r_l with the total response energy $\kappa_{v,v}^{(l)}$ to quantify editing capacity beyond the blind response.

The lower panels of Figure 2 show that ρ_l generally increases toward the output while r_l declines across both Qwen-VL backbones. Yet E_l forms non-monotonic *interaction ridges* around layers 21–24 in Qwen3-VL and 19–21 in Qwen2.5-VL-7B. These patterns suggest a depth-wise trade-off: shallow layers retain visual specificity but offer limited target control, whereas stronger control at depth increasingly overlaps with the blind response. Thus, E_l identifies layers where visual specificity and target controllability coexist, motivating selection by grounded editability rather than depth alone.

4 Methodology

4.1 Visual-Intervention Guided Residualization

Under the Fisher geometry, the local steepest-descent direction of the video-conditioned objective $\mathcal{L}_v$ is $-\boldsymbol{t}_{v,l}$, which may still contain a substantial blind component. VIGOR instead restricts the update to the descent orientation of the grounded interaction tangent $\boldsymbol{t}_l^{\mathrm{GI}}$ in Eq.(9). Since $\boldsymbol{g}_{v,l}^\top \boldsymbol{t}_l^{\mathrm{GI}} = E_l$ and $\boldsymbol{g}_{\varnothing,l}^\top \boldsymbol{t}_l^{\mathrm{GI}} = 0$, the direction $-\boldsymbol{t}_l^{\mathrm{GI}}$ decreases $\mathcal{L}_v$ while leaving $\mathcal{L}_\varnothing$ unchanged to first order. The resulting update thus targets what the video contributes beyond the text-side shortcut.

This residualization is a geometric projection rather than an auxiliary loss: it removes exactly the component aligned with the blind tangent, without introducing a tunable trade-off between editing effectiveness and suppression of the blind response. It turns the Visual-Null intervention from an output-level diagnostic into a constraint on the admissible editing direction.

4.2 Fisher-Geometric Minimum-Information Editing

Having isolated the grounded direction, we next determine how strongly and where to apply it. Let $d > 0$ denote the desired first-order reduction in $\mathcal{L}_v$. At layer l, VIGOR seeks the least functionally disruptive update that realizes this reduction while remaining insensitive to the Visual-Null pathway:

$$\min_{\Delta \boldsymbol{W}_l} \frac{1}{2}(\Delta \boldsymbol{W}_l)^\top \boldsymbol{F}_l (\Delta \boldsymbol{W}_l) \quad \text{s.t.} \quad \boldsymbol{g}_{v,l}^\top \Delta \boldsymbol{W}_l = -d, \qquad \boldsymbol{g}_{\varnothing,l}^\top \Delta \boldsymbol{W}_l = 0. \tag{11}$$

The objective minimizes the local distributional change defined by Eq.(7), thereby mitigating undesirable model drift. We can derive the following closed-form solution, with the complete derivation deferred to Appendix C:

$$\Delta \boldsymbol{W}_l^\star = -\frac{d}{E_l}\boldsymbol{t}_l^{\mathrm{GI}} = -\frac{d}{E_l}\boldsymbol{F}_l^{-1}\left(\boldsymbol{g}_{v,l} - \frac{\kappa_{v,\varnothing}^{(l)}}{\kappa_{\varnothing,\varnothing}^{(l)}}\boldsymbol{g}_{\varnothing,l}\right). \tag{12}$$

which reveals a direct role for the geometry developed in Section 3.2: $\boldsymbol{t}_l^{\mathrm{GI}}$ determines the grounded update direction, while E_l determines the scale required to achieve the target reduction d. Substituting Eq.(12) into the Fisher objective yields the minimum functional cost required at layer l:

$$J_l^\star \triangleq \frac{1}{2}(\Delta \boldsymbol{W}_l^\star)^\top \boldsymbol{F}_l (\Delta \boldsymbol{W}_l^\star) = \frac{d^2}{2E_l}. \tag{13}$$

For a fixed target reduction d, a larger E_l therefore implies a lower minimum functional cost, allowing the grounded edit to be realized with less disturbance to the model distribution. This relation also provides a principled criterion for layer localization:

$$l^\star = \arg\min_{l \in \mathcal{L}} J_l^\star = \arg\max_{l \in \mathcal{L}} E_l. \tag{14}$$

For an individual editing request, Eq.(14) identifies the layer achieving the prescribed local improvement at minimum Fisher cost. In practice, we select one fixed editing layer per backbone from the training-set E_l profile and reuse it across all evaluation requests. This selection follows the interaction ridges identified in Section 3.2, avoiding the large computational cost of per-request layer search. The boundedness and limiting behavior of $J_l^\star$ are provided in Appendix D.

5 Experiments

To comprehensively evaluate our VIGOR, we organize experiments around the following questions:

- **RQ1:** How does VIGOR perform on Video-LLM editing tasks compared to mainstream editors?
- **RQ2:** Can VIGOR mitigate *Visual Blindness* by suppressing edit-induced gains that remain accessible without instance-specific visual evidence?
- **RQ3:** How do VIGOR's core techniques and candidate editing layers affect its grounded editing performance, and do the observed trends align with the geometric principles underlying its design?
- **RQ4:** Can VIGOR preserve general video understanding capabilities while limiting representation drift under sequential editing?

5.1 Experiment Setup

We outline the datasets, metrics, backbones, and baselines below; more details appear in Appendix A.

Datasets & Metrics. We evaluate VIGOR on VMEB (Zhang et al., 2026), the first benchmark specifically designed for Video-LLM editing. Following VMEB, we report Reliability (*Rel.*), Text Locality (*Txt-Loc.*), Video Locality (*Vid-Loc.*), Generality (*Gen.*), and Robustness (*Rob.*). However, these outcome-level metrics cannot distinguish whether edit-induced improvements remain accessible after visual content is removed. We therefore further analyze the metrics defined in Section 3.

Backbones & Baseline Editors. We use four representative Video-LLM backbones spanning architectures, model generations, and scales: LLaVA-NeXT-Video (7B) (Zhang et al., 2024), Qwen2.5-VL (3B and 7B) (Bai et al., 2025b), and Qwen3-VL (8B) (Bai et al., 2025a). Following VMEB (Zhang et al., 2026), we compare against six editors adapted to Video-LLMs: *parameter-modifying* methods, including fine-tuning (FT) (Cheng et al., 2023), MEND (Mitchell et al., 2022a), MEMIT (Meng et al., 2023), and AlphaEdit (Fang et al., 2025); and *parameter-preserving* methods, including IKE (Zheng et al., 2023) and SERAC (Mitchell et al., 2022b). For FT, we evaluate language-only, vision-only, and joint vision–language variants, denoted as FT (Lang.), FT (Vis.), and FT (Vis-Lang.), respectively.

Implementation Details. For all baselines, we follow the official VMEB implementations and hyperparameters (Zhang et al., 2026). For VIGOR, we estimate the diagonally Kronecker-factored empirical Fisher once on the VMEB training split and reuse it across all editing requests. Each edit constructs its Visual-Null counterpart and applies Eq.(12) within a Fisher trust region. Based on the training-set E_l profile, we use one fixed FFN layer per backbone: 7 for LLaVA-NeXT-Video-7B, 8 for Qwen2.5-VL-3B, 19 for Qwen2.5-VL-7B, and 27 for Qwen3-VL-8B. The interaction ridges in Figure 2 are estimated on the balanced 100-example analysis set, whereas the deployed layers are selected from the full training-set E_l profile. Full implementation details, including Visual-Null construction, trust-region settings, and other hyperparameters, are provided in Appendix A.5.

5.2 Comprehensive Editing Performance Comparison (RQ1)

Table 1 shows that VIGOR achieves the highest average score on all backbones, combining perfect reliability (100.00%) with strong generality, robustness, and locality. FT attains high reliability but substantially degrades video locality, whereas MEMIT and AlphaEdit preserve video locality at the expense of editing success. VIGOR balances correction and preservation, retaining 98.68–99.71% text locality and 80.90–85.80% video locality. These results show that constraining the Visual-Null response is compatible with effective editing and behavioral preservation. Since these metrics do not reveal whether editing gains persist under Visual-Null, RQ2 examines this directly.

Table 1: Comparison of editors across four Video-LLM backbones on VMEB. *AVG.* is the arithmetic mean of the five metrics; Δ denotes VIGOR's average-score improvement.

Model	Editor	Editing Metrics					Overall	
		Rel.↑	T-Loc.↑	V-Loc.↑	Gen.↑	Rob.↑	AVG.↑	Δ
LLaVA-NeXT-Video Model size: 7B	FT (Lang.)	99.96	79.93	22.79	**99.92**	**99.96**	80.51	+16.03
	FT (Vis.)	99.95	**100.00**	13.48	98.46	99.74	82.33	+14.21
	FT (Vis-Lang.)	99.96	98.08	23.31	99.96	99.96	84.25	+12.29
	MEND	96.31	99.02	85.11	96.12	96.21	94.55	+1.99
	MEMIT	43.00	84.16	**98.77**	42.93	42.77	62.33	+34.21
	AlphaEdit	42.13	56.78	91.85	42.12	41.92	54.96	+41.58
	IKE	87.68	42.26	24.49	86.39	87.87	65.74	+30.80
	SERAC	80.23	99.87	96.39	79.51	78.98	87.00	+9.54
	VIGOR	**100.00**	99.71	84.49	99.56	98.96	**96.54**	0.00
Qwen2.5-VL Model size: 3B	FT (Lang.)	100.00	94.03	55.90	**99.84**	**100.00**	89.95	+5.71
	FT (Vis.)	99.79	**100.00**	32.58	98.98	91.93	84.66	+11.00
	FT (Vis-Lang.)	100.00	97.48	48.54	99.89	100.00	89.18	+6.49
	MEND	99.05	98.53	74.85	98.46	98.73	93.92	+1.74
	MEMIT	40.90	90.66	**99.02**	40.56	40.90	62.41	+33.25
	AlphaEdit	40.91	68.46	86.39	41.10	40.91	55.55	+40.11
	IKE	77.90	71.06	21.41	79.36	77.74	65.49	+30.17
	SERAC	83.27	99.35	89.75	82.71	80.17	87.05	+8.61
	VIGOR	**100.00**	99.19	83.31	97.16	98.67	**95.67**	0.00
Qwen2.5-VL Model size: 7B	FT (Lang.)	100.00	74.18	9.92	**99.81**	**99.95**	76.77	+17.60
	FT (Vis.)	99.56	**100.00**	30.77	97.50	89.36	83.44	+10.93
	FT (Vis-Lang.)	100.00	95.73	43.90	99.95	100.00	87.92	+6.45
	MEND	96.97	97.43	71.80	96.18	96.43	91.76	+2.61
	MEMIT	39.71	78.86	94.62	40.10	39.71	58.60	+35.77
	AlphaEdit	39.32	86.25	**97.39**	40.19	39.32	60.49	+33.88
	IKE	85.85	80.75	21.08	79.48	84.96	70.42	+23.95
	SERAC	75.46	99.77	88.09	75.57	75.06	82.79	+11.58
	VIGOR	**100.00**	98.68	80.90	95.81	96.45	**94.37**	0.00
Qwen3-VL Model size: 8B	FT (Lang.)	100.00	61.82	10.96	**99.84**	**99.84**	74.49	+20.85
	FT (Vis.)	99.65	**100.00**	29.14	98.15	90.45	83.48	+11.86
	FT (Vis-Lang.)	100.00	94.93	28.50	99.10	100.00	84.51	+10.83
	MEND	94.73	98.11	59.67	93.55	91.57	87.53	+7.81
	MEMIT	39.15	92.12	**96.15**	40.14	39.15	61.34	+34.00
	AlphaEdit	39.29	83.11	95.16	39.50	39.29	59.27	+36.07
	IKE	99.15	73.49	22.24	98.99	99.04	78.58	+16.76
	SERAC	82.13	99.46	87.48	81.07	80.89	86.21	+9.13
	VIGOR	**100.00**	99.51	85.80	98.10	93.29	**95.34**	0.00

Table 2: **Visual-Null dependence evaluation across four Video-LLMs.** VNull-Rel. denotes absolute post-edit target reliability under Visual-Null, whereas $\Delta_{\text{null}}/\Delta_{\text{full}}$ measures the edit-induced gain retained, with values closer to zero indicating weaker blind editing gains.

Editor	LLaVA-NeXT-Video		Qwen2.5-VL-3B		Qwen2.5-VL-7B		Qwen3-VL	
	VNull-Rel.↓	$\Delta_{\text{null}}/\Delta_{\text{full}}$ →0	VNull-Rel.↓	$\Delta_{\text{null}}/\Delta_{\text{full}}$ →0	VNull-Rel.↓	$\Delta_{\text{null}}/\Delta_{\text{full}}$ →0	VNull-Rel.↓	$\Delta_{\text{null}}/\Delta_{\text{full}}$ →0
FT(Lang.)	99.96	1.003	99.12	1.390	99.47	1.400	95.15	1.285
FT(Vis.)	79.40	0.222	**63.70**	**0.006**	**60.85**	0.005	76.83	0.048
FT(Vis-Lang.)	100.00	1.003	98.94	1.377	99.39	1.398	94.95	1.280
MEND	94.20	0.913	95.03	1.259	85.04	1.151	89.36	0.982
IKE	92.91	1.189	89.51	2.360	89.50	1.725	99.48	1.712
VIGOR	**73.27**	**-0.052**	66.16	-0.078	63.59	**0.002**	**75.55**	**-0.010**

5.3 Visual-Null Dependence Evaluation (RQ2)

We examine whether edit-induced gains persist after instance-specific visual evidence is removed. Table 2 shows that FT (Lang.) and FT (Vis-Lang.) retain 94.95–100.00% reliability under Visual-Null, with gain ratios exceeding one across all backbones, indicating that their editing gains remain largely accessible through the blind pathway. In contrast, VIGOR yields near-zero gain ratios (-0.078 to 0.002) while retaining near-perfect reliability with the original video, showing that its edit-induced improvements are largely suppressed under Visual-Null. FT (Vis.) provides an informative control: restricting updates to the vision module also suppresses blind gains. VIGOR achieves comparable suppression while editing only a single FFN layer in the language backbone, with substantially stronger overall editing performance, particularly in video locality. This suggests that suppressing blind editing gains does not require restricting updates to the vision module; it can also be achieved by constraining the update direction within the language backbone. We further examine request-level editing gains with and without instance-specific visual evidence across all

Table 3: **Ablations of Visual-Null residualization and Fisher geometry across four Video-LLMs.** We report mean ± standard deviation over three random seeds: 41, 42, and 43.

Variant	Standard Editing Metrics					Visual-Null Dependence	
	Rel.↑	T-Loc.↑	V-Loc.↑	Gen.↑	Rob.↑	VNull-Rel.↓	$\Delta_{\text{null}}/\Delta_{\text{full}}$ →0
LLaVA-NeXT-Video Model size: 7B							
VIGOR	**100.00 ± 0.00**	**99.71 ± 0.04**	**84.49 ± 0.02**	**99.56 ± 0.02**	**98.96 ± 0.02**	**73.26 ± 0.02**	**-0.052 ± 0.004**
w/o Visual-Null	99.96 ± 0.00	99.79 ±0.01	90.26 ± 0.04	99.90 ± 0.02	99.92 ± 0.01	99.81 ± 0.03	0.971 ± 0.0003
w/o Fisher	92.36 ± 0.02	99.78 ± 0.01	88.56 ± 0.01	89.20 ± 0.02	88.75 ± 0.00	74.80 ± 0.00	−0.0085 ± 0.001
Qwen2.5-VL Model size: 3B							
VIGOR	**100.00 ± 0.00**	**99.40 ± 0.15**	**83.39 ± 0.05**	**96.93 ± 0.16**	**98.62 ± 0.04**	**65.97 ± 0.13**	**-0.078 ± 0.000**
w/o Visual-Null	100.00 ± 0.00	99.53 ± 0.10	83.72 ± 0.03	98.26 ± 0.00	99.46 ± 0.00	92.82 ± 0.16	1.113 ± 0.001
w/o Fisher	95.43 ± 0.15	64.28 ± 0.75	26.85 ± 0.00	90.97 ± 0.10	91.23 ± 0.06	66.44 ± 0.44	0.024 ± 0.009
Qwen2.5-VL Model size: 7B							
VIGOR	**100.00 ± 0.00**	**99.37 ± 0.49**	**81.60 ± 0.51**	**95.85 ± 0.10**	**96.38 ± 0.05**	**63.60 ± 0.05**	**0.001 ± 0.001**
w/o Visual-Null	100.00 ± 0.00	99.72 ± 0.03	81.88 ± 0.11	96.33 ± 0.05	97.45 ± 0.07	75.50 ± 0.19	0.982 ± 0.001
w/o Fisher	88.85 ± 0.00	99.74 ± 0.00	82.19 ± 0.00	80.78 ± 0.00	82.78 ± 0.00	61.68 ± 0.00	−0.002 ± 0.000
Qwen3-VL Model size: 8B							
VIGOR	**100.00 ± 0.00**	**99.29 ± 0.16**	**85.28 ± 0.36**	**97.90 ± 0.14**	**93.76 ± 0.34**	**75.44 ± 0.08**	**0.006 ± 0.011**
w/o Visual-Null	100.00 ± 0.00	99.53 ± 0.05	85.30 ± 0.08	98.26 ± 0.06	94.84 ± 0.09	85.72 ± 0.15	0.751 ± 0.002
w/o Fisher	91.71 ± 0.06	99.56 ± 0.04	86.14 ± 0.07	90.14 ± 0.09	85.98 ± 0.11	74.88 ± 0.12	−0.012 ± 0.004

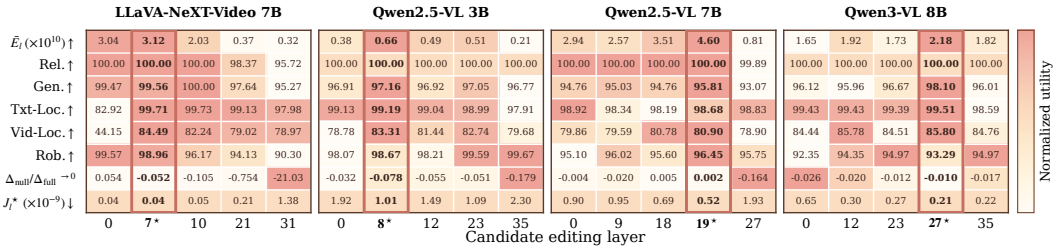


Figure 3: **Effect of editing location across four Video-LLMs.** We compare editing performance, Visual-Null gain ratios, $\bar{E}_l$, and observed Fisher functional cost $J_l^\star$ across candidate layers. Outlined columns and stars mark selected layers. Detailed numerical results are provided in Table 4.

backbones (Figure 7, Appendix B.2). The same qualitative behavior is observed when VIGOR is instantiated with alternative Visual-Null constructions, including an unrelated video and a repeated static frame (Appendix B.5).

5.4 Geometric Ablations and Layer-wise Analysis (RQ3)

Ablation Study. We evaluate two variants to isolate VIGOR's core components: *w/o Visual-Null* removes the residualization constraint while retaining Fisher geometry; *w/o Fisher* replaces Fisher geometry with Euclidean geometry while retaining Visual-Null residualization. Table 3 shows that residualization and Fisher geometry serve complementary roles. Without residualization, reliability remains near-perfect, but Visual-Null gain ratios rise to 0.751–1.113, revealing blind gains masked by conventional editing success. Without Fisher geometry, these ratios remain near zero, but reliability drops by 4.57–11.15 percentage points, generality and robustness decline, and video locality on Qwen2.5-VL-3B falls from 83.39% to 26.85%. Residualization thus suppresses blind gains, while Fisher geometry supports effective correction with less functional disruption; together, they achieve 100% reliability across all four backbones with near-zero Visual-Null gain ratios.

Effect of Editing Location. Figure 3 examines how editing location affects grounded editability and functional cost. Across all backbones, the selected layer simultaneously maximizes the aggregated grounded editability $\bar{E}_l$ and minimizes the observed Fisher functional cost $J_l^\star$, providing empirical support for the inverse relationship derived at the request level in Section 3.2. Reliability remains near saturation across most candidate layers, indicating that editing success alone provides little guidance for localization. In contrast, $\bar{E}_l$ distinguishes layers where target controllability coexists with a response beyond the Visual-Null pathway, and its maximizer supports the desired correction at the lowest functional cost. The selected layers also maintain strong generality, locality, and robustness, showing that the geometric criterion remains predictive beyond the local first-order analysis.

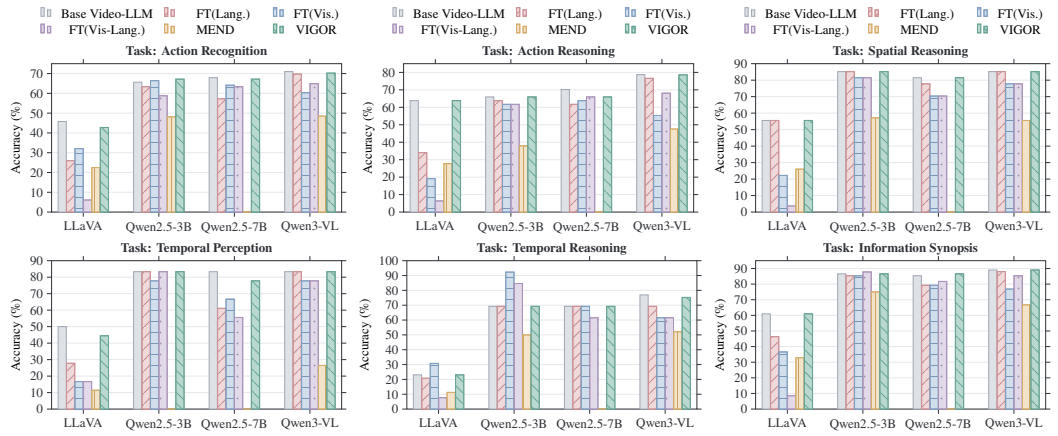


Figure 4: **General capability preservation after 100 sequential edits.** Accuracy on six representative tasks from the Video-MME short-video subset, with unedited base models as references.

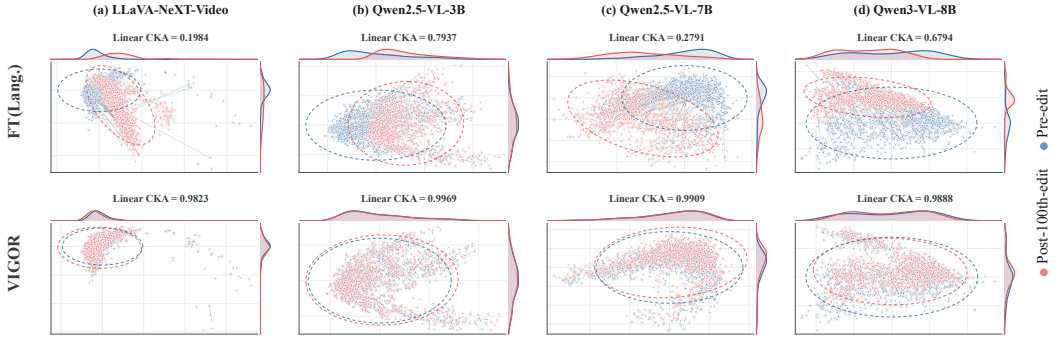


Figure 5: **Hidden-state distributions before and after 100 sequential edits with linear CKA scores.**

5.5 General Capability Preservation and Representation Stability (RQ4)

We evaluate general capabilities on the Video-MME short-video subset and hidden-state stability after 100 sequential edits to each backbone. Experimental details appear in Appendix B.4.

General Capability Preservation. Figure 4 compares editors across six representative tasks spanning visual perception, spatiotemporal reasoning, and global video understanding. Results for all 12 tasks appear in Figure 8, with further analysis in Appendix B.4. After 100 sequential edits, VIGOR remains close to the unedited models across most backbone–task combinations, whereas other editors exhibit more pronounced, task-dependent degradation.

Representation Stability. The hidden-state distributions in Figure 5 reveal complementary evidence of preservation at the representation level. VIGOR maintains linear CKA scores of 0.9823–0.9969 across all four backbones, compared with 0.1984–0.7937 for FT (Lang.), indicating substantially greater structural similarity to the unedited representations. Together, these results suggest that VIGOR limits cumulative drift in both model behavior and internal representations. This finding is consistent with VIGOR's minimum-information objective, which penalizes local changes in the model's predictive distribution through Fisher geometry.

6 Conclusion

Video-LLM editing can succeed without genuinely relying on visual evidence, exposing a gap between behavioral correction and video-grounding. We bridge this gap with Visual-Null intervention and Fisher geometry, which identify and remove blind editing responses. This leads to VIGOR, a Fisher-orthogonal minimum-information editor that targets the video-conditioned response beyond the blind pathway. Across four Video-LLMs, VIGOR achieves effective editing with strongly suppressed Visual-Null gains while preserving general capabilities. Our findings highlight video-grounding as a distinct dimension of Video-LLM editing beyond output correction alone.

AI USE STATEMENT

In this work, generative AI tools were used solely for language and grammar polishing during the writing process. They were not used in the design of the research, experimental analysis, interpretation of the results, or development of the core contributions of this work. All AI-assisted text was carefully reviewed and verified by the authors. The authors take full responsibility for all aspects of this work, including the manuscript, research claims, and any artifacts produced with the assistance of AI.

REPRODUCIBILITY STATEMENT

We are committed to promoting the transparency and reproducibility of this work. The complete experimental code is provided in the appendix, along with detailed instructions to facilitate independent verification and replication of our results.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Qizhou Chen, Chengyu Wang, Dakan Wang, Taolin Zhang, Wangyue Li, and Xiaofeng He. Lifelong knowledge editing for vision language models with low-rank mixture-of-experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 9455–9466, 2025a.

Qizhou Chen, Taolin Zhang, Chengyu Wang, Xiaofeng He, Dakan Wang, and Tingting Liu. Attribution analysis meets model editing: Advancing knowledge correction in vision language models with visedit. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2168–2176, 2025b.

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 13877–13888, 2023.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5484–5495, 2021.

Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. VLKEB: A large vision-language model knowledge editing benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Kailin Jiang, Yuntao Du, Yukai Ding, Yuchen Ren, Ning Jiang, Zhi Gao, Zilong Zheng, Lei Liu, Bin Li, and Qing Li. When large multimodal models confront evolving knowledge: Challenges and explorations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025.

Siyuan Li, Youyuan Zhang, Fangming Liu, and Jing Li. Modality-decoupled online recursive editing. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=4VMbVE64lF.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL http://jmlr.org/papers/v21/17-678.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Proceedings of the Advances in neural information processing systems (NeurIPS)*, pp. 17359–17372, 2022.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022a.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 15817–15831, 2022b.

Zhiyi Shi, Binjie Wang, Chongjie Si, Yichen Wu, Junsik Kim, and Hanspeter Pfister. Dualedit: Dual editing for knowledge updating in vision-language models. In *Proceedings of the Second Conference on Language Modeling (COLM)*, 2025.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Chenghao Xu, Jiexi Yan, Guangtao Lyu, Qi Liu, Muli Yang, and Cheng Deng. Fisher-driven adaptive locating for knowledge editing in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20902–20913, 2026.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

Qian Zhang, Xinye Li, Xiaokai Wu, Junhao Xu, Zhanyue Qin, Qingbin Liu, Junxian Cai, Xi Chen, Bolin Zhang, Zhiying Tu, et al. Editing the moving world: Model editing for video llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6072–6091, 2026.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4862–4876, 2023.

APPENDIX

A Experimental Setup **14**

A.1 Datasets . . . 14

A.2 Evaluation Metrics . . . 14

A.2.1 Editing Effectiveness . . . 14

A.2.2 Dependence under Visual-Null . . . 15

A.2.3 General Capability Preservation . . . 15

A.2.4 Representation Stability . . . 15

A.3 Video-LLM Backbones . . . 15

A.4 Baseline Editors . . . 16

A.5 Implementation Details . . . 16

B Supplementary Experiments **17**

B.1 Extended Analysis of Editing Attribution via Visual Intervention . . . 17

B.2 Visual-Null Dependence Evaluation (RQ2) . . . 18

B.3 Geometric Ablations and Layer-wise Analysis (RQ3) . . . 19

B.4 General Capability Preservation and Representation Stability (RQ4) . . . 20

B.5 Robustness to Visual-Null Construction . . . 22

B.6 Case Study . . . 23

C Closed-Form Solution of Fisher-Geometric Editing **28**

D Boundedness and Limiting Behavior of $J_l^\star$ **29**

A EXPERIMENTAL SETUP

In this section, we provide detailed descriptions of experimental setup, including explanation of datasets and evaluation metrics, discussion of baseline methods and implementation details.

A.1 DATASETS

We use two complementary benchmarks: VMEB for evaluating editing effectiveness and dependence under Visual-Null, and Video-MME for assessing general video understanding.

- **VMEB** (Zhang et al., 2026): VMEB contains 1,578 editing instances across 14 video subcategories drawn from MVBench (Li et al., 2024) and Video-MME (Fu et al., 2025). Each instance consists of a video–question pair with an annotated target answer, together with evaluation inputs for text/video locality, question-rephrasing generality, and robustness to video perturbations. We follow the official training and evaluation split. For VIGOR, the training split is used as reference data for empirical Fisher estimation; the resulting matrices are computed once per backbone and reused across editing requests, with all evaluation instances excluded from Fisher estimation. To assess whether editing gains persist without instance-specific visual content, we additionally construct Visual-Null counterparts of the editing inputs by removing instance-specific visual information while preserving the questions, target answers, and multimodal input structure.
- **Video-MME** (Fu et al., 2025): Video-MME comprises 900 videos and 2,700 manually annotated multiple-choice questions spanning six domains and three duration groups. We use its short-video subset, containing videos shorter than two minutes, and evaluate accuracy across 12 task categories. Comparing each backbone before editing and after 100 sequential edits measures the preservation of general video understanding beyond request-level locality. Since VMEB partly draws from Video-MME, we remove overlapping examples between the editing requests and the Video-MME evaluation set, ensuring that capability preservation is assessed on examples not used for editing.

A.2 EVALUATION METRICS

We evaluate editing effectiveness, dependence under Visual-Null, and preservation of general capabilities and representations. Let $e = (x^v, x^q, y^\star) \sim \mathcal{D}_{\text{edit}}$ denote an editing request, $T_e = |y^\star|$ its target length, and $\theta_{\Delta W}$ the corresponding edited parameters.

A.2.1 EDITING EFFECTIVENESS

We report five metrics averaged over editing requests. Reliability, generality, and robustness measure target-token accuracy, while locality measures prediction agreement before and after editing. All scores are reported as percentages; higher is better.

Reliability (Rel.). Reliability measures target-answer accuracy on the original video–question pair.

$$\text{Rel.} = \mathbb{E}_{e\sim\mathcal{D}_{\text{edit}}}\left[\frac{1}{T_e}\sum_{t=1}^{T_e}\mathbf{1}\left\{\arg\max_w p_{\theta_{\Delta W}}(w \mid x^v, x^q, y^\star_{<t}) = y^\star_t\right\}\right]. \tag{15}$$

Generality (Gen.). Generality measures target-answer accuracy under semantically equivalent question rephrasings $\widetilde{x}^q \sim \mathcal{R}(x^q)$, with the video and target answer unchanged:

$$\text{Gen.} = \mathbb{E}_{e\sim\mathcal{D}_{\text{edit}}}\mathbb{E}_{\widetilde{x}^q\sim\mathcal{R}(x^q)}\left[\frac{1}{T_e}\sum_{t=1}^{T_e}\mathbf{1}\left\{\arg\max_w p_{\theta_{\Delta W}}(w \mid x^v, \widetilde{x}^q, y^\star_{<t}) = y^\star_t\right\}\right]. \tag{16}$$

Robustness (Rob.). Robustness measures target-answer accuracy under semantics-preserving video perturbations $\widetilde{x}^v \sim \mathcal{P}(x^v)$, such as rotation and playback-speed changes, with the question and target answer unchanged:

$$\text{Rob.} = \mathbb{E}_{e\sim\mathcal{D}_{\text{edit}}}\mathbb{E}_{\widetilde{x}^v\sim\mathcal{P}(x^v)}\left[\frac{1}{T_e}\sum_{t=1}^{T_e}\mathbf{1}\left\{\arg\max_w p_{\theta_{\Delta W}}(w \mid \widetilde{x}^v, x^q, y^\star_{<t}) = y^\star_t\right\}\right]. \tag{17}$$

Text Locality (T-Loc.). Text locality measures prediction agreement before and after editing on unrelated text-only inputs from $\mathcal{U}_e^{\text{text}}$. For each locality input, let y^{loc} denote its reference sequence and $T_{\text{loc}} = |y^{\text{loc}}|$. Both models are evaluated under the same reference prefixes:

$$\text{T-Loc.} = \mathbb{E}_{e\sim\mathcal{D}_{\text{edit}}}\mathbb{E}_{x^q\sim\mathcal{U}_e^{\text{text}}}\left[\frac{1}{T_{\text{loc}}}\sum_{t=1}^{T_{\text{loc}}}\mathbf{1}\left\{\begin{array}{c}\arg\max_w p_{\theta_{\Delta W}}(w \mid x^q, y^{\text{loc}}_{<t}) \\ = \arg\max_w p_\theta(w \mid x^q, y^{\text{loc}}_{<t})\end{array}\right\}\right]. \tag{18}$$

Video Locality (V-Loc.). Video locality measures prediction agreement on unrelated video–question pairs from $\mathcal{U}_e^{\text{video}}$, using the same reference-prefix convention:

$$\text{V-Loc.} = \mathbb{E}_{e\sim\mathcal{D}_{\text{edit}}}\mathbb{E}_{(x^v,x^q)\sim\mathcal{U}_e^{\text{video}}}\left[\frac{1}{T_{\text{loc}}}\sum_{t=1}^{T_{\text{loc}}}\mathbf{1}\left\{\begin{array}{c}\arg\max_w p_{\theta_{\Delta W}}(w \mid x^v, x^q, y^{\text{loc}}_{<t}) \\ = \arg\max_w p_\theta(w \mid x^v, x^q, y^{\text{loc}}_{<t})\end{array}\right\}\right]. \tag{19}$$

A.2.2 Dependence under Visual-Null

We evaluate editing outcomes under Visual-Null using two complementary metrics:

- **Visual-Null Reliability (VNull-Rel.).** Post-edit reliability on $(x^{v,\varnothing}, x^q)$, using the original target and the same teacher-forced token-level scoring procedure. This measures absolute target-token accuracy without informative video, including predictive ability already present before editing.
- **Editing Gain Ratio ($\Delta_{\text{null}}/\Delta_{\text{full}}$).** The fraction of original-video editing gain retained under Visual-Null, computed from the loss reductions defined in the main text. For an evaluation set $\mathcal{D} = \{e_i\}_{i=1}^N$, we report

$$R_{\text{gain}}(\mathcal{D}) = \frac{\sum_{i=1}^N \Delta_{\text{null},i}}{\sum_{i=1}^N \Delta_{\text{full},i}} = \frac{\overline{\Delta}_{\text{null}}}{\overline{\Delta}_{\text{full}}}. \tag{20}$$

  Aggregating before division avoids the instability of averaging per-request ratios with near-zero denominators. For positive aggregate original-video gain, ratios near zero indicate negligible net Visual-Null gain, while ratios near one indicate comparable gains under both inputs. Negative ratios indicate that editing increases the Visual-Null target loss on average.

VNull-Rel. measures absolute post-edit performance, whereas the gain ratio isolates edit-induced improvement. Both are interpreted alongside original-video reliability: suppressing blind gains is meaningful only if the requested correction remains effective.

A.2.3 General Capability Preservation

We report multiple-choice accuracy across the 12 task categories of the Video-MME short-video subset before editing and after 100 sequential edits. Performance relative to the unedited model measures the cumulative effect of editing on general video understanding.

A.2.4 Representation Stability

We compare corresponding hidden representations before editing and after 100 sequential edits using three complementary metrics:

- **Linear CKA** measures similarity between the original and edited representation structures; higher values indicate greater structural preservation.
- **Cosine Drift** defined as one minus the cosine similarity between corresponding hidden vectors, measuring directional change. Lower values indicate better preservation.
- **Relative ℓ_2 Drift** measures the Euclidean distance between corresponding hidden vectors normalized by the original vector's norm.

A.3 Video-LLM Backbones

We evaluate VIGOR on four Video-LLM backbones spanning different architectures, model generations, and scales: LLaVA-NeXT-Video (7B), Qwen2.5-VL (3B and 7B), and Qwen3-VL (8B). All backbones are evaluated under the official VMEB settings (Zhang et al., 2026).

- **LLaVA-NeXT-Video (7B).** LLaVA-NeXT-Video (Zhang et al., 2024) extends the LLaVA-NeXT framework to video understanding. It encodes sampled video frames with a visual encoder and maps their representations into the language embedding space through a multimodal projector. The language backbone then processes the visual representations together with the textual query to generate a response. We use the 7B variant as a representative LLaVA-based Video-LLM.
- **Qwen2.5-VL (3B and 7B).** Qwen2.5-VL (Bai et al., 2025b) supports image and video understanding through a visual encoder coupled with a language backbone. Its visual processing supports dynamic spatial resolution and frame sampling, while temporal information is incorporated through its positional encoding. We evaluate both the 3B and 7B variants to examine editing behavior across model scales within the same family.
- **Qwen3-VL (8B).** Qwen3-VL (Bai et al., 2025a) further develops the Qwen vision–language architecture with DeepStack integration of visual features and interleaved multimodal rotary position embeddings. Its video representation incorporates temporal information through explicit timestamp alignment. We use the 8B variant to assess whether the Visual-Null analysis and editing method extend to a newer generation of Video-LLMs.

A.4 Baseline Editors

Fine-tuning (FT). FT directly optimizes the target response over selected parameters while freezing the rest (Cheng et al., 2023). We evaluate three variants: FT (Lang.), FT (Vis.), and FT (Vis-Lang.), which update language-side parameters, vision-side parameters, and both, respectively.

MEND. Model Editor Networks with Decomposition (MEND) (Mitchell et al., 2022a) is a learned parameter editor. It uses lightweight hypernetworks to transform the low-rank factors of edit-loss gradients and construct updates to the selected weights. After editor-specific training, these hypernetworks generate updates for new editing requests. Following VMEB (Zhang et al., 2026), we evaluate language-side variants by applying the editor to the corresponding parameter subsets.

MEMIT. Mass-Editing Memory in a Transformer (MEMIT) (Meng et al., 2023) builds on the FFN key–value view of factual associations. It optimizes target representations for the requested edits and distributes the corresponding weight updates across multiple language FFN layers. Activation statistics are incorporated to balance fitting the new associations against preserving existing mappings. We adopt its VMEB adaptation (Zhang et al., 2026) for video-conditioned editing under the edit–evaluate–restore protocol.

AlphaEdit. AlphaEdit (Fang et al., 2025) introduces null-space constraints into locate-then-edit methods. It constructs a projection from the activation keys of preservation examples and constrains weight updates to reduce changes to the protected mappings. We adopt its VMEB adaptation (Zhang et al., 2026) as a preservation-constrained parameter editor. Unlike VIGOR's constraint on the Visual-Null response, AlphaEdit's constraint is defined by the preservation keys.

IKE. In-Context Knowledge Editing (IKE) (Zheng et al., 2023) incorporates edited knowledge through in-context learning without modifying model parameters. It retrieves relevant demonstrations and combines them with the target fact to construct an augmented prompt that guides generation.

SERAC. Semi-parametric Editing with a Retrieval-Augmented Counterfactual (SERAC) (Mitchell et al., 2022b) stores edits in an external memory while keeping the backbone parameters frozen. A learned scope classifier determines whether an input is related to a stored edit. Inputs within the edit scope are handled by a counterfactual model, while other inputs are answered by the original model. We use its VMEB adaptation as a memory-based alternative to direct parameter modification.

A.5 Implementation Details

Editable parameters and layer selection. We implement VIGOR on top of the official VMEB codebase,[3] and use the official implementations and hyperparameters for all baselines (Zhang et al., 2026). VIGOR requires neither an auxiliary editor nor an external optimizer. For each request,

[3] https://github.com/Sakabamrisa/VMEB

it updates only the FFN output-projection matrix $\boldsymbol{W}_l$ at a single language-backbone layer, while keeping all other parameters frozen. Based on the layer-wise geometric profiles computed on the VMEB training split, we select one fixed editing layer per backbone and reuse it across all evaluation requests. Using zero-based indexing, the selected layers are 7 for LLaVA-NeXT-Video (7B), 8 for Qwen2.5-VL (3B), 19 for Qwen2.5-VL (7B), and 27 for Qwen3-VL (8B). All remaining VIGOR hyperparameters are shared across backbones.

Reference Fisher geometry. For each backbone, we estimate a block-diagonal, diagonally Kronecker-factored empirical Fisher using only the VMEB training split; no evaluation instance is used. Each block factorizes the per-feature second moments of the FFN inputs and output gradients. The resulting statistics are computed once and reused across all editing requests, separating offline geometry estimation from request-time editing. We use relative damping of 10^{-3} with a numerical floor of 10^{-12} when inverting the factors. The editable matrix, gradients, and Fisher computations are performed in FP32, while the remainder of the backbone retains its native precision.

Paired objectives and Visual-Null construction. For each request $(x^v, x^q, y^\star)$, the video-conditioned and Visual-Null objectives share the same question, target sequence, teacher-forced target prefixes, and mean target-token negative log-likelihood. We construct $x^{v,\varnothing}$ after video preprocessing by replacing only `pixel_values_videos` with an all-zero tensor of identical shape, dtype, and device. All structural inputs—including visual placeholders, video-grid and temporal metadata, position indices, attention masks, and the visual-token budget—remain unchanged. Thus, the paired objectives differ only in the availability of instance-specific visual content rather than in the multimodal input structure. Visual-Null should therefore be viewed as a controlled intervention rather than a naturalistic blank video.

Iterative closed-form editing. At editing step k, we compute the video-conditioned and Visual-Null gradients at the current parameters $\theta^{(k)}$, while keeping the precomputed Fisher fixed. The requested reduction in video-conditioned loss is

$$d^{(k)} = \max\left\{\mathcal{L}_v(\theta^{(k)}; x^v, x^q) - \ell_{\text{tar}},\, 0\right\}, \qquad \ell_{\text{tar}} = 0.01. \tag{21}$$

Equivalently, this targets a mean target-token log-probability of -0.01. We apply Eq.(12) with a per-step Fisher trust-region budget of 10^{-3}. If the target remains unmet, both gradients are recomputed and the local geometry is relinearized, for at most five closed-form steps. Editing terminates early once $\mathcal{L}_v(\theta^{(k)}; x^v, x^q) \le \ell_{\text{tar}}$; requests already satisfying this condition receive no update.

Evaluation and reproducibility. For independent-edit evaluation, we follow an edit–evaluate–restore protocol, restoring the original model weights before processing the next request. For sequential-edit evaluation, we instead apply 100 edits without restoration and evaluate Video-MME only after the final edit, using the unedited backbone as the reference. Unless otherwise specified, all experiments use random seed 42 and are run on a single server with eight NVIDIA L40S GPUs. The supplementary materials include source code, hyperparameter configurations, precomputed Fisher factors, and additional experimental results.

B Supplementary Experiments

B.1 Extended Analysis of Editing Attribution via Visual Intervention

Experimental Setup. We construct a balanced analysis set of 100 examples from the VMEB training split, randomly sampling 20 examples from each of five source groups: CLEVRER, Data, FunQA, STAR, and Validation. All examples are strictly disjoint from the test split; "Validation" denotes a source-group label within the training data. For editing attribution, we use FT (Lang.) as the representative conventional editor and compare the gains induced by each update on the original input and its structure-preserving Visual-Null counterpart. For mechanism analysis, we characterize layer-wise responses using Video–Null Fisher alignment ρ_l, grounded fraction r_l, and video-grounded editability E_l, as defined in the main text.

Figure 6 presents supplementary results for Qwen2.5-VL-3B and LLaVA-NeXT-Video-7B. Both models exhibit substantial blind editing gains, yet differ in where visual specificity and target controllability coexist across depth.

Blind Responses across Architectures. On Qwen2.5-VL-3B, 64 of 100 edits satisfy $\Delta_{\text{null}} \geq \Delta_{\text{full}}$, and the aggregate Visual-Null gain reaches $1.11\times$ the original-video gain. On LLaVA-NeXT-Video-7B, 48 of 100 edits satisfy this condition, while Visual-Null retains 96.9% of the original-video gain in aggregate. These results extend the main-text findings across model scales and architectures: answer-level editing success can largely survive the removal of instance-specific visual evidence.

Architecture-Dependent Editing Geometry. The lower panels in Figure 6 reveal a common depth-wise trade-off with architecture-specific structure. For Qwen2.5-VL-3B, grounded editability forms a primary ridge around layers 8–10 and a weaker secondary ridge around layers 25–30. Although shallow layers retain a large grounded fraction r_l, their editability remains limited, indicating that visual specificity alone does not provide sufficient control over the target response. At greater depth, the video-conditioned and Visual-Null responses become increasingly aligned under Fisher geometry, progressively shrinking the component that remains after removing the blind response.

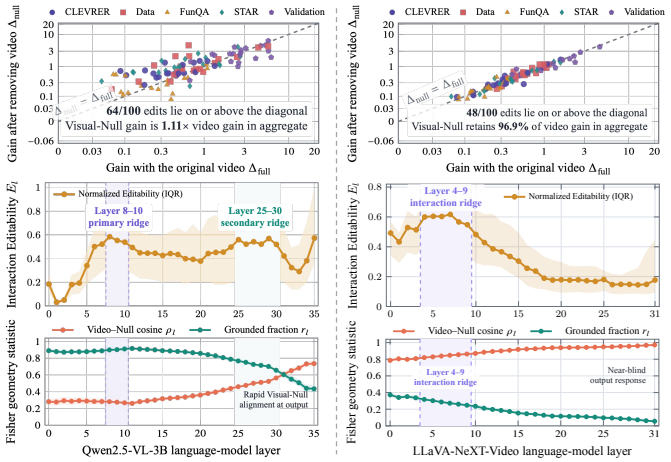


Figure 6: **Cross-architecture analysis of editing attribution and response geometry.** Top: editing gains with the original video and under Visual-Null; the dashed diagonal denotes equal gains. Middle: normalized video-grounded editability E_l across layers, with shaded interquartile ranges. Bottom: Video–Null Fisher alignment ρ_l and grounded fraction $r_l = 1 - \rho_l^2$.

LLaVA-NeXT-Video-7B exhibits a different profile, with a pronounced editability ridge across layers 4–9 and layer 7 attaining the highest average grounded editability used for layer selection. Beyond this region, editability declines as the video-conditioned response becomes increasingly aligned with the Visual-Null pathway. Thus, while the location and shape of the interaction ridge vary across architectures, both backbones exhibit the same underlying tension between visual specificity and target controllability. These observations motivate locating edits by grounded editability E_l rather than by network depth alone: effective editing occurs where sufficient target control coexists with a response component that cannot be explained by the Visual-Null pathway.

B.2 Visual-Null Dependence Evaluation (RQ2)

Figure 7 extends the main-text analysis to four backbones, comparing VIGOR with FT (Lang.), FT (Vis.), and FT (Vis-Lang.). Each point represents an editing request, contrasting its gain with the original video, Δ_{full}, against that under Visual-Null, Δ_{null}. These distributions reveal whether successful corrections remain effective after informative visual evidence is removed.

Blind Vision Gains across Backbones. FT (Lang.) exhibits substantial blind gains on all four backbones, with distinct distributional patterns. On LLaVA-NeXT-Video, edits closely follow the equal-gain diagonal, and the aggregate Visual-Null gain matches the original-video gain. On Qwen2.5-VL-7B, Qwen2.5-VL-3B, and Qwen3-VL, respectively, 343, 357, and 319 of 472 edits lie on or above the diagonal; the corresponding aggregate gain ratios are 1.40, 1.39, and 1.28. Thus, informative video is unnecessary for much of the improvement induced by language-side fine-tuning. FT (Vis-Lang.) produces similar distributions, showing that jointly updating visual and language modules does not resolve this ambiguity in the editing objective.

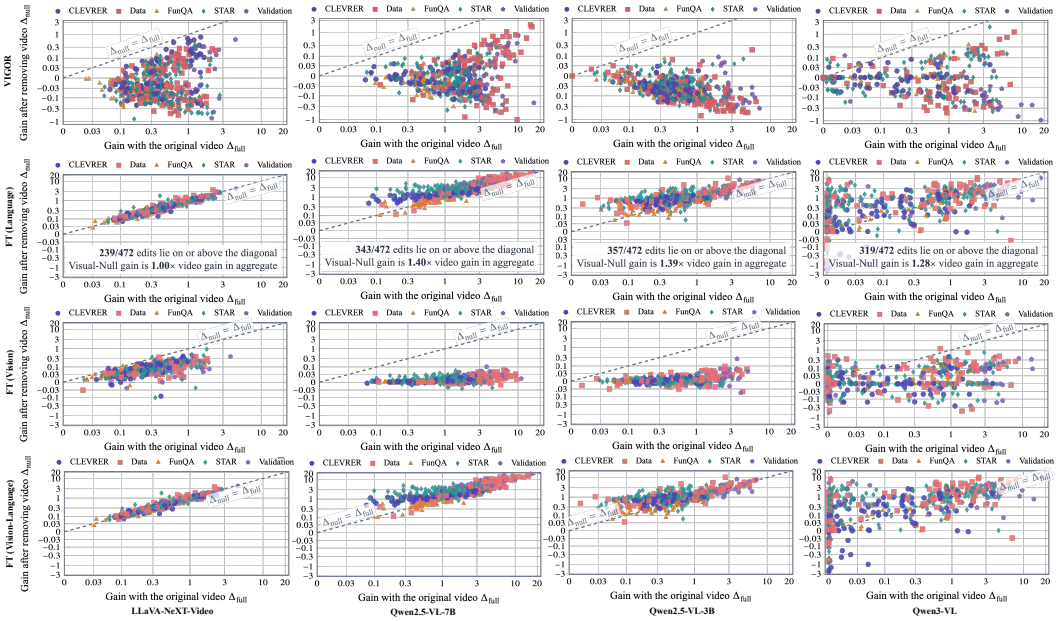


Figure 7: **Extended visual-Null dependence evaluation of VIGOR and all FT variants across four Video-LLMs.** VIGOR suppresses blind gains while retaining video-conditioned improvements.

Suppressing Blind Gains beyond Module Selection. FT (Vis.) substantially suppresses Visual-Null gains, particularly on the two Qwen2.5-VL backbones. However, directly modifying the visual encoder can perturb shared visual representations and affect behavior beyond the editing target, consistent with its weaker video locality in the main-text results (Table 1). Most parameter-modifying baselines evaluated here instead operate within the language backbone, motivating constraints that suppress blind editing gains within these layers. VIGOR achieves this through Fisher-orthogonal residualization in a single language-backbone FFN layer. Across all four backbones, its editing gains largely separate from the equal-gain diagonal, combining positive Δ_{full} with small or negative Δ_{null}. On Qwen2.5-VL-3B, many edits lie below $\Delta_{\text{null}} = 0$: the target answer becomes more likely with the original video but less likely under Visual-Null. Together with its strong overall editing performance, these results show that blind gains under Visual-Null can be suppressed by constraining the update direction within the language backbone.

Absolute Reliability versus Edit-Induced Gains. The nonzero VNull-Rel. of VIGOR and FT (Vis.) in Table 2 (Section 5.3) does not necessarily indicate substantial blind gains. VNull-Rel. measures absolute post-edit token accuracy, which may reflect pre-existing language priors and target tokens predictable from the ground-truth prefix supplied during teacher forcing Meng et al. (2022). Therefore, we must distinguish between the model's pre-existing predictive ability and the improvement introduced by an edit. VIGOR constrains the latter under Visual-Null, so its near-zero gain ratios are compatible with absolute VNull-Rel. above 60%.

B.3 Geometric Ablations and Layer-wise Analysis (RQ3)

Table 4 provides the detailed numerical results underlying the main-text layer-wise heatmap. Across all four backbones, the geometry-selected layer attains the highest average grounded editability $\bar{E}_l$ and the lowest Fisher functional cost among the evaluated candidates, while maintaining near-perfect reliability and near-zero Visual-Null gain ratios. This pattern is consistent with the theoretical relation $J_l^\star = d^2/(2E_l)$ for a fixed target improvement: greater grounded editability permits the desired correction with lower functional cost.

Quantitatively, $\bar{E}_l$ and $J_l^\star$ exhibit strong inverse rank correlations across candidate layers, with Spearman coefficients of -1.00, -1.00, -1.00, and -0.70 for LLaVA-NeXT-Video, Qwen2.5-VL-3B, Qwen2.5-VL-7B, and Qwen3-VL, respectively. Reliability alone reveals little of this distinction, as it remains near saturation across layers with substantially different grounded editability and

Table 4: Effect of candidate editing layers on VIGOR. For a backbone with L decoder layers, we evaluate four uniformly spaced reference layers, $l_q = \mathrm{round}(q(L-1))$ for $q \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$, together with the geometry-selected layer $l^\star = \arg\max_l \bar{E}_l$.

Backbone	Layer	$\bar{E}_l \uparrow$	Rel.↑	Gen.↑	Txt-Loc.↑	Vid-Loc.↑	Rob.↑	$\Delta_{\mathrm{null}}/\Delta_{\mathrm{full}} \to 0$	Cost $J_l^\star \downarrow$
LLaVA-NeXT Video (Model size: 7B)	0	3.039×10^{10}	100.00	99.47	82.92	44.15	99.57	0.054	4.059×10^{-11}
	7 ($l^\star$)	3.121×10^{10}	100.00	99.56	99.71	84.49	98.96	-0.052	4.034×10^{-11}
	10	2.028×10^{10}	100.00	100.00	99.73	82.24	96.17	-0.105	5.433×10^{-11}
	21	3.736×10^{9}	98.37	97.64	99.13	79.02	94.13	-0.754	2.100×10^{-10}
	31	3.241×10^{9}	95.72	95.27	97.98	78.97	90.30	-21.030	1.381×10^{-9}
Qwen2.5-VL (Model size: 3B)	0	3.825×10^{9}	100.00	96.91	99.13	78.78	98.07	-0.032	1.922×10^{-9}
	8 ($l^\star$)	6.607×10^{9}	100.00	97.16	99.19	83.31	98.67	-0.078	1.006×10^{-9}
	12	4.904×10^{9}	100.00	96.92	99.04	81.44	98.21	-0.055	1.486×10^{-9}
	23	5.124×10^{9}	100.00	97.05	98.99	82.74	99.59	-0.051	1.094×10^{-9}
	35	2.147×10^{9}	100.00	96.77	97.91	79.68	99.67	-0.179	2.302×10^{-9}
Qwen2.5-VL (Model size: 7B)	0	2.935×10^{10}	100.00	94.76	98.92	79.86	95.10	-0.004	8.961×10^{-10}
	9	2.572×10^{10}	100.00	95.03	98.34	79.59	96.02	-0.020	9.530×10^{-10}
	18	3.514×10^{10}	100.00	94.76	98.19	80.78	95.60	0.005	6.885×10^{-10}
	19 ($l^\star$)	4.601×10^{10}	100.00	95.81	98.68	80.90	96.45	0.002	5.175×10^{-10}
	27	8.090×10^{9}	99.89	93.07	98.83	78.90	95.75	-0.164	1.928×10^{-9}
Qwen3-VL (Model size: 8B)	0	1.652×10^{10}	100.00	96.12	99.43	84.44	92.35	-0.026	6.514×10^{-10}
	12	1.924×10^{10}	100.00	95.96	99.43	85.78	94.35	-0.020	3.002×10^{-10}
	23	1.733×10^{10}	100.00	96.67	99.39	84.51	94.97	-0.012	2.672×10^{-10}
	27 ($l^\star$)	2.178×10^{10}	100.00	98.10	99.51	85.80	93.29	-0.010	2.083×10^{-10}
	35	1.819×10^{10}	100.00	96.01	98.59	84.76	94.97	-0.017	2.249×10^{-10}

functional costs. Together, these results show that the geometric criterion provides a consistent signal for edit localization across architectures, while the selected layers also maintain strong generality, locality, and robustness under finite updates.

B.4 General Capability Preservation and Representation Stability (RQ4)

Task Coverage and Evaluation Setup. We evaluate the short-video subset of Video-MME across 12 task categories, covering complementary aspects of video understanding:

- **Recognition and perception:** Action Recognition, Object Recognition, Attribute Perception, Spatial Perception, and Temporal Perception assess the identification of actions, entities, attributes, and spatial or temporal information.
- **Reasoning and quantification:** Action Reasoning, Object Reasoning, Spatial Reasoning, Temporal Reasoning, and Counting assess inference and quantification over video content.
- **Text and global understanding:** OCR evaluates text recognition within videos, while Information Synopsis assesses understanding of their central information.

Following the main-text protocol, we evaluate each backbone after 100 sequential edits, using its unedited performance as the reference. Overlapping examples between editing requests and the evaluation set are removed. For representation analysis, we use the same Video-MME short-video evaluation set and collect paired hidden states before editing and after 100 sequential VIGOR edits. For each video–question input, we extract the output of the final language-model normalization layer at the last prompt token immediately before answer generation. For visualization, the paired pre- and post-edit representations are jointly projected to two dimensions using PCA.

General Capability Preservation. Figure 8 shows that VIGOR largely preserves general video understanding after 100 sequential edits, remaining close to the unedited models across most backbone–task combinations. This preservation spans recognition, reasoning, counting, and text understanding, suggesting that VIGOR suppresses the functional drift associated with cumulative editing interference. The contrast is particularly pronounced on LLaVA-NeXT-Video, where several baselines degrade across perception and reasoning, while VIGOR largely retains the original capability profile. The Qwen backbones exhibit a similar pattern of broad capability retention, with task-dependent variation.

MEND presents a contrasting failure on Qwen2.5-VL-7B: after 100 sequential edits, accuracy falls to zero across all 12 evaluated tasks, indicating model collapse under accumulated updates. This result exposes the gap between isolated edit success and sequential stability: individually effective

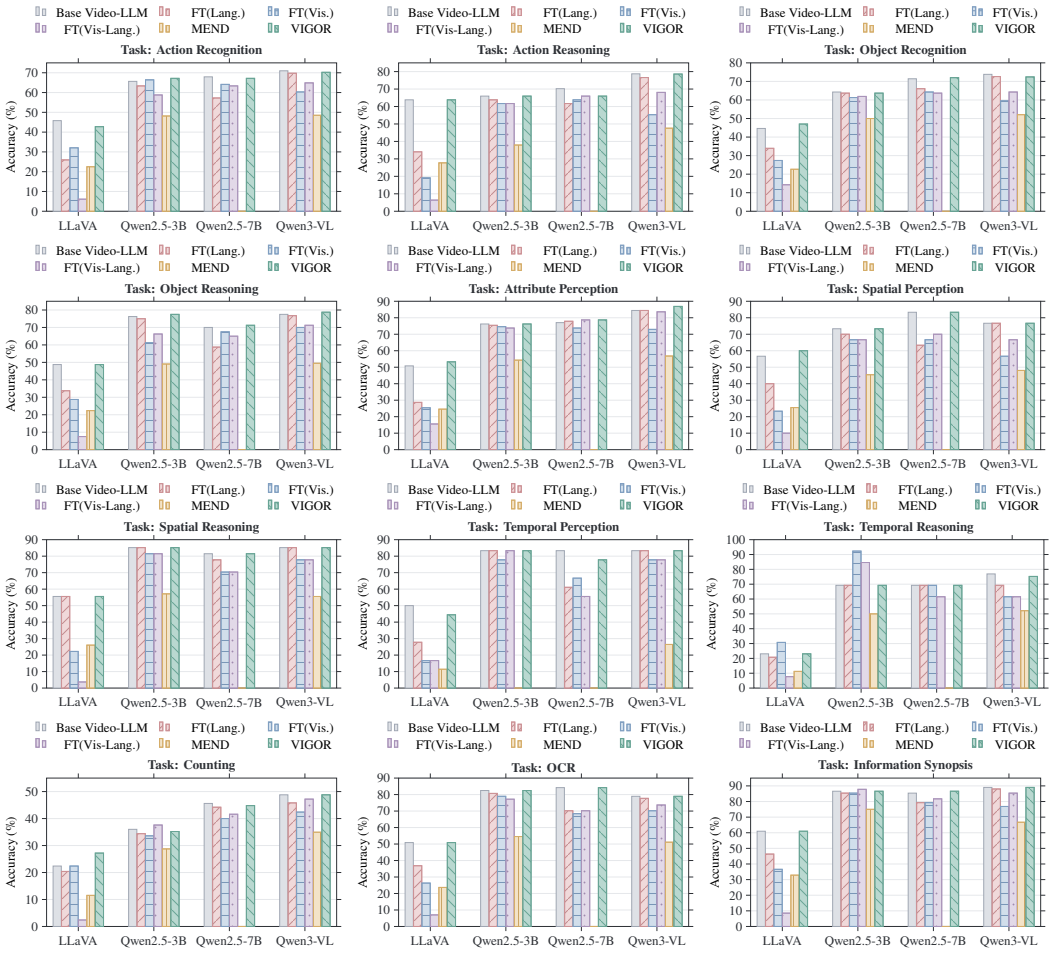


Figure 8: **General capability preservation after 100 sequential edits.** Accuracy across 12 Video-MME tasks and four backbones, with unedited models as references. VIGOR largely preserves the model's general-purpose visual capabilities.

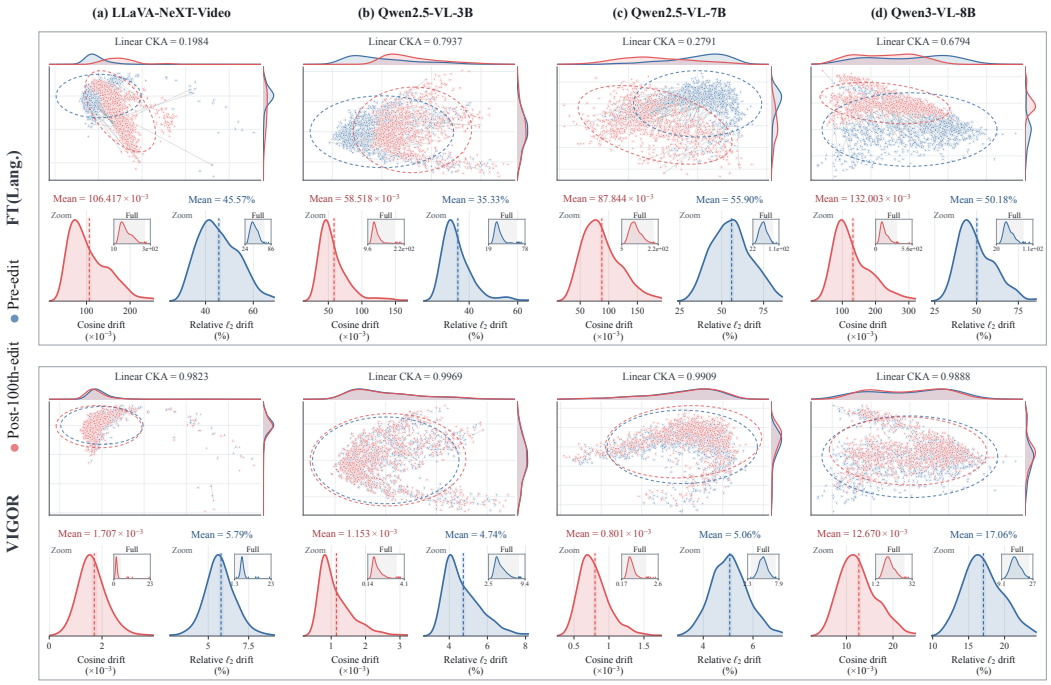


Figure 9: **Detailed representation stability analysis after 100 sequential edits.** Top: hidden-state distributions before and after editing with linear CKA scores. Bottom: cosine drift and relative ℓ_2 drift distributions measuring directional and magnitude changes of representations.

corrections can collectively disrupt capabilities beyond the editing targets. VIGOR's sustained performance on the same backbone supports its ability to contain this cumulative interference.

These findings connect targeted editing to the preservation of broader model functionality. Consistent with its minimum-information design, VIGOR limits functional drift as edits accumulate, allowing local corrections while largely retaining general video understanding.

Representation Drift. Beyond the hidden-state visualization and linear CKA analysis in the main text, we further quantify representation changes after 100 sequential edits using cosine drift and relative ℓ_2 drift. Cosine drift measures directional deviation, while relative ℓ_2 drift captures the magnitude of displacement relative to the original representation.

Figure 9 provides a detailed comparison between VIGOR and FT (Lang.). Across all four backbones, VIGOR consistently induces smaller representation changes. Its mean cosine drift ranges from $(0.801\text{–}12.670)\times10^{-3}$, compared with $(58.518\text{–}132.003)\times10^{-3}$ for FT (Lang.). Similarly, VIGOR achieves lower mean relative ℓ_2 drift (4.74–17.06%) than FT (Lang.) (35.33–55.90%). Together with the linear CKA results reported in the main text, these measurements show that VIGOR better preserves representation structure under sequential editing. This stability aligns with our Fisher-geometric minimum-information objective, which reduces unnecessary functional perturbations while achieving the desired edit.

B.5 Robustness to Visual-Null Construction

The effectiveness of VIGOR relies on the Visual-Null intervention to identify editing responses that bypass informative visual evidence. A potential concern is whether the observed suppression of blind editing gains depends on a particular null-video realization, especially the default all-zero input. To examine this, we evaluate VIGOR with two alternative Visual-Null constructions that attenuate different components of the informative visual signal.

Specifically, we introduce *Cross-Video Null*, where the original video is replaced by an unrelated video from the multimodal-locality split of the same VMEB example, while keeping the textual query and target answer unchanged. The substituted video is processed by the original video pipeline,

Table 5: **Robustness to Visual-Null construction.** We evaluate VIGOR with three null interventions: the default Zero Null, Cross-Video Null, and Static-Frame Null. All results are obtained with random seed 42. $\Delta_{\text{null}}/\Delta_{\text{full}}$ measures the edit-induced gain retained under the corresponding Visual-Null intervention.

Null Construction	Standard Editing Metrics					Visual-Null Dependence	
	Rel.$^{\uparrow}$	T-Loc.$^{\uparrow}$	V-Loc.$^{\uparrow}$	Gen.$^{\uparrow}$	Rob.$^{\uparrow}$	VNull-Rel.$^{\downarrow}$	$\Delta_{\text{null}}/\Delta_{\text{full}}$ $^{\to 0}$
LLaVA-NeXT-Video Model size: 7B							
Zero Null (default)	100.00	99.71	84.49	99.56	98.96	73.27	−0.052
Cross-Video Null	99.78	99.66	78.79	99.35	97.63	71.88	−0.248
Static-Frame Null	99.05	99.50	72.60	97.84	96.01	72.24	−0.385
Qwen2.5-VL Model size: 3B							
Zero Null (default)	100.00	99.19	83.31	97.16	98.67	66.16	−0.078
Cross-Video Null	100.00	99.03	82.73	96.41	96.87	67.07	−0.171
Static-Frame Null	100.00	99.00	82.33	96.20	95.51	69.27	−0.231
Qwen2.5-VL Model size: 7B							
Zero Null (default)	100.00	98.68	80.90	95.81	96.45	63.59	0.002
Cross-Video Null	100.00	98.68	80.96	95.33	94.73	64.09	−0.015
Static-Frame Null	100.00	98.66	80.83	94.22	93.11	64.38	−0.027
Qwen3-VL Model size: 8B							
Zero Null (default)	100.00	99.51	85.80	98.10	93.29	75.55	−0.010
Cross-Video Null	100.00	98.87	84.11	97.26	92.08	76.31	−0.021
Static-Frame Null	99.79	98.74	83.42	96.68	90.62	77.56	−0.054

preserving consistent visual tokens, spatial metadata, and temporal metadata. We further introduce *Static-Frame Null*, which repeats the middle decoded frame of the original video across all temporal positions before preprocessing. This preserves static appearance, resolution, and frame budget while removing temporal variation. Therefore, Cross-Video Null removes instance-specific visual evidence while maintaining natural video statistics, whereas Static-Frame Null suppresses temporal and motion-dependent evidence while retaining static appearance cues.

For both interventions, we keep the editing requests, Fisher metric, candidate layers, and optimization hyperparameters unchanged. We only recompute the Visual-Null response used for grounded residualization. Table 5 shows that VIGOR retains high editing reliability and suppresses positive gains under the corresponding null interventions. However, preservation depends on the intervention, with a larger video-locality drop on LLaVA-NeXT-Video under Static-Frame Null. These results indicate that VIGOR is applicable beyond the default all-zero construction, while the choice of intervention affects the correction–preservation trade-off.

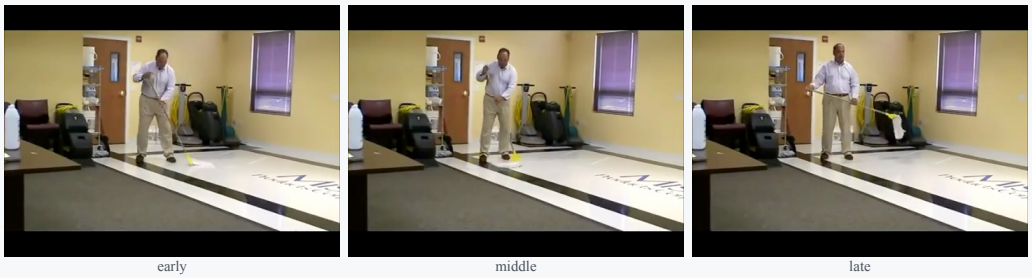
B.6 Case Study

To provide request-level evidence beyond the aggregate results, we visualize representative edits across all four Video-LLM backbones in Figures 10–13. The examples span human actions, semantic roles and physical processes, and controlled object motion. For each request, we report the model behavior before and after editing under both the original video and its Visual-Null counterpart. We compare VIGOR with FT (Lang.), which achieves strong conventional editing reliability but exhibits substantial Visual-Null gains in our aggregate evaluation.

Analysis. The cases expose a distinction that conventional editing success alone cannot capture. Across all 12 examples, both FT (Lang.) and VIGOR attain a post-edit reliability of 1.00 with the original video, with comparable Δ_{full}. Their interventional behavior is nevertheless markedly different. FT (Lang.) consistently produces large positive Visual-Null gains ($\Delta_{\text{null}} = 0.569$–$9.104$), often driving the target response to full reliability even after informative visual evidence is removed. In contrast, VIGOR keeps Δ_{null} close to zero (-0.173–0.061), so the edit contributes little additional preference for the target under Visual-Null.

These examples illustrate the central failure mode behind Visual Blindness: an editor can appear perfectly successful by strengthening a target association that remains retrievable without the video. VIGOR instead preserves the correction attributable to the original video-conditioned response while suppressing the component shared with the blind response. The request-level behavior therefore mirrors the aggregate results in RQ2 and provides an intuitive realization of VIGOR's geometric objective: the edit changes what the video contributes to the prediction, rather than merely making the desired answer easier to retrieve.

Case A: Human action recognition LLaVA-NeXT-Video-7B



early middle late

What activity does the video depict?

VMEB erroneous answer: washing ⟶ **Edit target:** mopping

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{full}	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{null}
FT (Lang.)	0.50	1.00	+0.850	0.50	1.00	+0.861
VIGOR	**0.50**	**1.00**	**+0.831**	**0.50**	**0.50**	**−0.037**

Case B: Role recognition LLaVA-NeXT-Video-7B



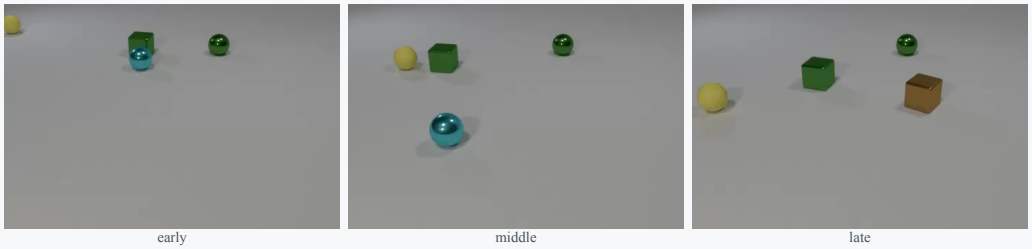
early middle late

What is the role of the man speaking in the video?

VMEB erroneous answer: eye pencil designer ⟶ **Edit target:** makeup artist

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{full}	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{null}
FT (Lang.)	0.67	1.00	+0.311	0.67	1.00	+0.569
VIGOR	**0.67**	**1.00**	**+0.302**	**0.67**	**0.67**	**+0.002**

Case C: Controlled motion LLaVA-NeXT-Video-7B



early middle late

In which direction does the green sphere move?

VMEB erroneous answer: up and to the right ⟶ **Edit target:** the object is stationary

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{full}	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{null}
FT (Lang.)	0.80	1.00	+1.363	0.80	1.00	+1.329
VIGOR	**0.80**	**1.00**	**+1.337**	**0.80**	**0.60**	**−0.173**

Figure 10: **Case studies on LLaVA-NeXT-Video-7B.** Both FT (Lang.) and VIGOR successfully correct the target response with the original video. Under Visual-Null, however, FT (Lang.) retains substantial editing gains, whereas VIGOR largely suppresses these gains.

Case A: Human action recognition Qwen2.5-VL-3B



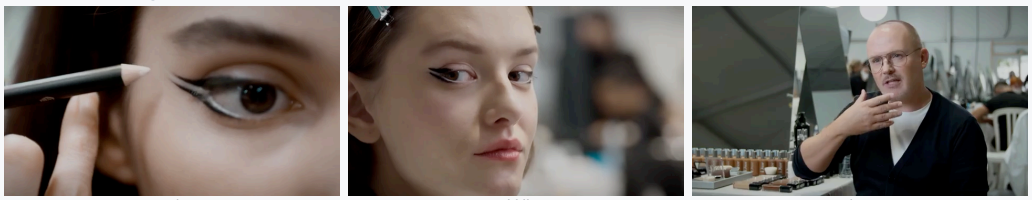
early middle late

What activity does the video depict?

VMEB erroneous answer: washing ⟶ **Edit target:** mopping

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{full}	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{null}
FT (Lang.)	0.50	1.00	+2.732	0.50	1.00	+3.249
VIGOR	**0.50**	**1.00**	**+2.665**	**0.50**	**0.50**	**−0.027**

Case B: Role recognition Qwen2.5-VL-3B



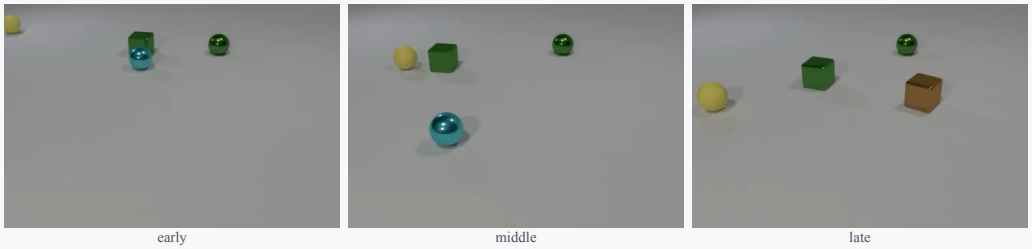
early middle late

What is the role of the man speaking in the video?

VMEB erroneous answer: eye pencil designer ⟶ **Edit target:** makeup artist

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{full}	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{null}
FT (Lang.)	0.67	1.00	+1.249	0.67	1.00	+2.771
VIGOR	**0.67**	**1.00**	**+1.234**	**0.67**	**0.67**	**−0.004**

Case C: Controlled motion Qwen2.5-VL-3B



early middle late

In which direction does the green sphere move?

VMEB erroneous answer: up and to the right ⟶ **Edit target:** the object is stationary

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{full}	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{null}
FT (Lang.)	0.75	1.00	+0.744	0.75	1.00	+1.310
VIGOR	**0.75**	**1.00**	**+0.723**	**0.75**	**0.75**	**−0.002**

Figure 11: **Case studies on Qwen2.5-VL-3B.** Across action recognition, role recognition, and controlled motion, VIGOR achieves the desired correction on the original video while suppressing the corresponding edit-induced gain after informative visual evidence is removed.

Case A: Human action recognition Qwen2.5-VL-7B



early middle late

What activity does the video depict?

VMEB erroneous answer: washing ⟶ **Edit target:** mopping

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{full}	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{null}
FT (Lang.)	0.50	1.00	+2.257	0.50	1.00	+5.325
VIGOR	**0.50**	**1.00**	**+2.201**	**0.50**	**0.50**	**+0.061**

Case B: Physical process recognition Qwen2.5-VL-7B



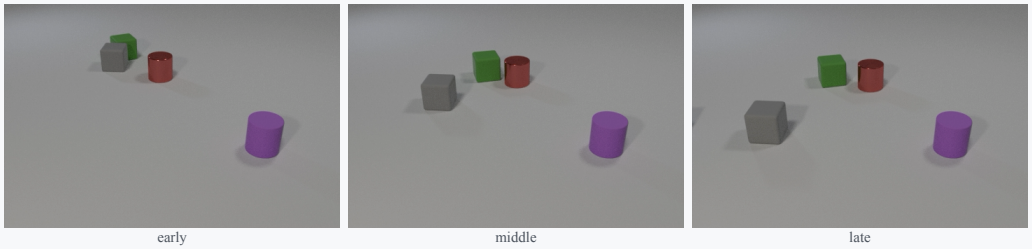
early middle late

Which description correctly matches the action?

VMEB erroneous answer: descending ⟶ **Edit target:** pouring

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{full}	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{null}
FT (Lang.)	0.50	1.00	+3.394	0.50	1.00	+9.104
VIGOR	**0.50**	**1.00**	**+3.424**	**0.50**	**0.50**	**0.000**

Case C: Controlled motion Qwen2.5-VL-7B



early middle late

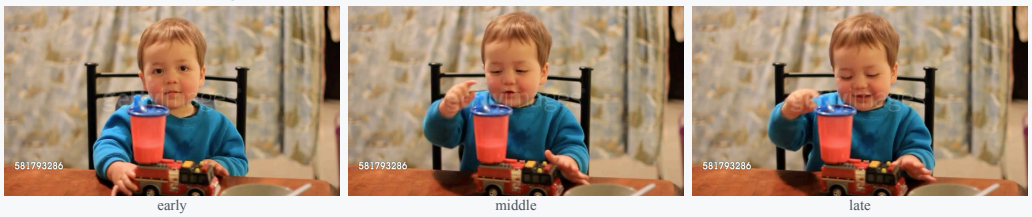
In which direction does the green cube move?

VMEB erroneous answer: up and to the right ⟶ **Edit target:** down and to the right

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{full}	**Pre-edit Rel.**	**Post-edit Rel.**	Δ_{null}
FT (Lang.)	0.80	1.00	+3.063	0.80	1.00	+2.888
VIGOR	**0.80**	**1.00**	**+3.031**	**0.80**	**0.80**	**−0.009**

Figure 12: **Case studies on Qwen2.5-VL-7B.** FT (Lang.) strongly increases the target likelihood under both the original video and Visual-Null, while VIGOR preserves the original-video correction with little change to the Visual-Null response.

Case A: Human action recognition Qwen3-VL-8B



early middle late

Which description correctly matches the action?

VMEB erroneous answer: feeding ⟶ **Edit target:** balancing

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	**Δ_{full}**	**Pre-edit Rel.**	**Post-edit Rel.**	**Δ_{null}**
FT (Lang.)	0.50	1.00	+3.671	0.50	1.00	+4.721
VIGOR	**0.50**	**1.00**	**+3.692**	**0.50**	**0.50**	**−0.025**

Case B: Multi-person motion Qwen3-VL-8B



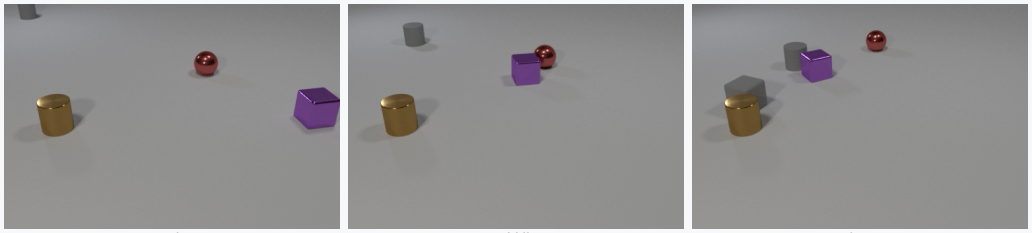
early middle late

How do the participants move during the competition?

VMEB erroneous answer: lie flat ⟶ **Edit target:** plank

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	**Δ_{full}**	**Pre-edit Rel.**	**Post-edit Rel.**	**Δ_{null}**
FT (Lang.)	0.50	1.00	+1.671	0.50	1.00	+3.645
VIGOR	**0.50**	**1.00**	**+1.720**	**0.50**	**0.50**	**−0.024**

Case C: Controlled motion Qwen3-VL-8B



early middle late

In which direction does the red sphere move?

VMEB erroneous answer: down and to the left ⟶ **Edit target:** up and to the left

	Original video			Visual-Null		
Editor	**Pre-edit Rel.**	**Post-edit Rel.**	**Δ_{full}**	**Pre-edit Rel.**	**Post-edit Rel.**	**Δ_{null}**
FT (Lang.)	0.80	1.00	+0.835	0.80	1.00	+2.800
VIGOR	**0.80**	**1.00**	**+0.776**	**0.80**	**0.80**	**+0.000**

Figure 13: **Case studies on Qwen3-VL-8B.** The same behavior generalizes to action recognition, multi-person motion, and controlled motion: VIGOR realizes the target correction when the relevant video is present without making the correction broadly accessible through the blind pathway.

C Closed-Form Solution of Fisher-Geometric Editing

We provide the derivation of the closed-form solution in Eq.(12). For notational convenience, parameter blocks and their gradients are implicitly identified with their vectorized coordinates whenever they appear in Fisher quadratic forms, inner products, or linear systems. We consider the non-degenerate case $\boldsymbol{F}_l \succ 0$, $\kappa^{(l)}_{\varnothing,\varnothing} > 0$, and $E_l > 0$.

Proposition 1 (Closed-form VIGOR update)**.** *For the constrained optimization problem*

$$\min_{\Delta \boldsymbol{W}_l} \frac{1}{2}(\Delta \boldsymbol{W}_l)^\top \boldsymbol{F}_l (\Delta \boldsymbol{W}_l) \quad \text{s.t.} \quad \boldsymbol{g}_{v,l}^\top \Delta \boldsymbol{W}_l = -d, \qquad \boldsymbol{g}_{\varnothing,l}^\top \Delta \boldsymbol{W}_l = 0, \tag{22}$$

the unique optimal update is

$$\Delta \boldsymbol{W}_l^\star = -\frac{d}{E_l} \boldsymbol{F}_l^{-1} \left(\boldsymbol{g}_{v,l} - \frac{\kappa^{(l)}_{v,\varnothing}}{\kappa^{(l)}_{\varnothing,\varnothing}} \boldsymbol{g}_{\varnothing,l} \right) = -\frac{d}{E_l} \boldsymbol{t}_l^{\mathrm{GI}}. \tag{23}$$

Proof. Since $\boldsymbol{F}_l \succ 0$, the objective in Eq.(22) is strictly convex, while both constraints are affine. Hence, the Karush–Kuhn–Tucker conditions are sufficient for global optimality.

The Lagrangian is

$$\mathscr{L}(\Delta \boldsymbol{W}_l, \lambda, \mu) = \frac{1}{2}(\Delta \boldsymbol{W}_l)^\top \boldsymbol{F}_l (\Delta \boldsymbol{W}_l) + \lambda \left(\boldsymbol{g}_{v,l}^\top \Delta \boldsymbol{W}_l + d \right) + \mu \, \boldsymbol{g}_{\varnothing,l}^\top \Delta \boldsymbol{W}_l. \tag{24}$$

Stationarity with respect to $\Delta \boldsymbol{W}_l$ gives

$$\boldsymbol{F}_l \Delta \boldsymbol{W}_l + \lambda \boldsymbol{g}_{v,l} + \mu \boldsymbol{g}_{\varnothing,l} = 0, \tag{25}$$

and therefore

$$\Delta \boldsymbol{W}_l = -\boldsymbol{F}_l^{-1} \left(\lambda \boldsymbol{g}_{v,l} + \mu \boldsymbol{g}_{\varnothing,l} \right). \tag{26}$$

Substituting Eq.(26) into the two constraints yields

$$\begin{bmatrix} \kappa^{(l)}_{v,v} & \kappa^{(l)}_{v,\varnothing} \\ \kappa^{(l)}_{v,\varnothing} & \kappa^{(l)}_{\varnothing,\varnothing} \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} d \\ 0 \end{bmatrix}. \tag{27}$$

Using

$$E_l = \kappa^{(l)}_{v,v} - \frac{\left(\kappa^{(l)}_{v,\varnothing} \right)^2}{\kappa^{(l)}_{\varnothing,\varnothing}}, \tag{28}$$

the determinant of the coefficient matrix becomes

$$\kappa^{(l)}_{v,v} \kappa^{(l)}_{\varnothing,\varnothing} - \left(\kappa^{(l)}_{v,\varnothing} \right)^2 = \kappa^{(l)}_{\varnothing,\varnothing} E_l > 0. \tag{29}$$

Thus, the KKT system has a unique solution. From its second row,

$$\mu = -\lambda \frac{\kappa^{(l)}_{v,\varnothing}}{\kappa^{(l)}_{\varnothing,\varnothing}}, \tag{30}$$

which, after substitution into the first row, gives

$$\lambda = \frac{d}{E_l}. \tag{31}$$

Substituting λ and μ into Eq.(26) yields

$$\begin{aligned} \Delta \boldsymbol{W}_l^\star &= -\frac{d}{E_l} \boldsymbol{F}_l^{-1} \left(\boldsymbol{g}_{v,l} - \frac{\kappa^{(l)}_{v,\varnothing}}{\kappa^{(l)}_{\varnothing,\varnothing}} \boldsymbol{g}_{\varnothing,l} \right) \\ &= -\frac{d}{E_l} \boldsymbol{t}_l^{\mathrm{GI}}, \end{aligned} \tag{32}$$

which proves Eq.(23). □

The solution also satisfies

$$\boldsymbol{g}_{v,l}^{\top}\boldsymbol{t}_l^{\mathrm{GI}} = E_l, \qquad \boldsymbol{g}_{\varnothing,l}^{\top}\boldsymbol{t}_l^{\mathrm{GI}} = 0, \tag{33}$$

making explicit that $\boldsymbol{t}_l^{\mathrm{GI}}$ determines the feasible grounded direction, while E_l determines the scale required to achieve the prescribed first-order reduction d.

Degenerate Visual-Null response. If $\kappa_{\varnothing,\varnothing}^{(l)} = 0$, positive definiteness of $\boldsymbol{F}_l^{-1}$ implies $\boldsymbol{g}_{\varnothing,l} = 0$. In this case, the Visual-Null constraint is vacuous and no blind component needs to be removed; the optimization reduces to the single-constraint Fisher minimum-norm update.

D Boundedness and Limiting Behavior of $J_l^\star$

We next analyze the minimum functional cost $J_l^\star$ associated with the closed-form update. The following result characterizes its dependence on grounded editability.

Proposition 2 (Functional-cost properties)**.** *For a fixed target reduction $d > 0$ and any layer satisfying $E_l > 0$, the minimum Fisher functional cost is*

$$J_l^\star = \frac{d^2}{2E_l}. \tag{34}$$

Consequently, $J_l^\star$ is strictly decreasing in E_l, remains bounded whenever E_l is bounded away from zero, and diverges as $E_l \to 0^+$.

Proof. By the definition of the grounded interaction tangent,

$$\begin{aligned} \left\|\boldsymbol{t}_l^{\mathrm{GI}}\right\|_{\boldsymbol{F}_l}^2 &= \left(\boldsymbol{g}_{v,l} - \frac{\kappa_{v,\varnothing}^{(l)}}{\kappa_{\varnothing,\varnothing}^{(l)}}\boldsymbol{g}_{\varnothing,l}\right)^{\top} \boldsymbol{F}_l^{-1} \left(\boldsymbol{g}_{v,l} - \frac{\kappa_{v,\varnothing}^{(l)}}{\kappa_{\varnothing,\varnothing}^{(l)}}\boldsymbol{g}_{\varnothing,l}\right) \\ &= \kappa_{v,v}^{(l)} - \frac{\left(\kappa_{v,\varnothing}^{(l)}\right)^2}{\kappa_{\varnothing,\varnothing}^{(l)}} = E_l. \end{aligned} \tag{35}$$

Substituting $\Delta\boldsymbol{W}_l^\star = -(d/E_l)\boldsymbol{t}_l^{\mathrm{GI}}$ into the Fisher objective gives

$$\begin{aligned} J_l^\star &= \frac{1}{2}(\Delta\boldsymbol{W}_l^\star)^{\top}\boldsymbol{F}_l(\Delta\boldsymbol{W}_l^\star) \\ &= \frac{d^2}{2E_l^2}\left\|\boldsymbol{t}_l^{\mathrm{GI}}\right\|_{\boldsymbol{F}_l}^2 = \frac{d^2}{2E_l}, \end{aligned} \tag{36}$$

which proves Eq.(34).

For fixed d, its derivative with respect to E_l is

$$\frac{\partial J_l^\star}{\partial E_l} = -\frac{d^2}{2E_l^2} < 0, \tag{37}$$

so the minimum functional cost decreases monotonically as grounded editability increases.

If $E_l \geq \epsilon > 0$, then

$$J_l^\star = \frac{d^2}{2E_l} \leq \frac{d^2}{2\epsilon}, \tag{38}$$

establishing boundedness whenever grounded editability is bounded away from zero. Conversely,

$$\lim_{E_l \to 0^+} J_l^\star = +\infty. \tag{39}$$

Thus, achieving the same grounded editing effect becomes increasingly expensive as the visually specific editable component vanishes. □

The limiting case $E_l = 0$ also admits a geometric interpretation. By the Cauchy–Schwarz inequality under the $\boldsymbol{F}_l^{-1}$ metric,

$$\left(\kappa_{v,\varnothing}^{(l)}\right)^2 \leq \kappa_{v,v}^{(l)} \kappa_{\varnothing,\varnothing}^{(l)}, \tag{40}$$

and hence $E_l \geq 0$. Equality $E_l = 0$ holds exactly when the Video and Visual-Null gradients are linearly dependent in the corresponding Fisher geometry. In this case, any update satisfying

$$\boldsymbol{g}_{\varnothing,l}^{\top} \Delta \boldsymbol{W}_l = 0 \tag{41}$$

necessarily satisfies

$$\boldsymbol{g}_{v,l}^{\top} \Delta \boldsymbol{W}_l = 0, \tag{42}$$

and therefore cannot simultaneously satisfy $\boldsymbol{g}_{v,l}^{\top} \Delta \boldsymbol{W}_l = -d$ for $d > 0$. Thus, a non-trivial first-order grounded edit is infeasible when $E_l = 0$.

Finally, for a fixed target reduction d, Eq.(34) immediately implies

$$\arg\min_{l \in \mathcal{L}} J_l^{\star} = \arg\max_{l \in \mathcal{L}} E_l, \tag{43}$$

which formally justifies the request-level layer-selection criterion underlying VIGOR. In practice, to avoid the substantial computational cost of per-request layer search, VIGOR approximates this criterion by selecting a fixed editing layer from the aggregated training-set editability profile and reusing it across evaluation requests. Since Eq.(7) is a local second-order approximation, these guarantees apply to the Fisher functional cost; the corresponding KL interpretation holds locally within the validity regime of the Fisher approximation.