MONITOR MORE, SUPERVISE LESS: CROSS-MODAL SAFETY MONITORING FOR OMNI-MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-models can receive harmful requests through different modalities and their combinations. Yet safety monitors that detect such requests from internal representations often generalize poorly beyond the modalities on which they are trained. With n modalities, a request can appear in up to $2^n - 1$ nonempty modality configurations, making exhaustive safety supervision and continual updating costly. This raises a central question: Can we achieve effective cross-modal safety monitoring in a data-efficient way? We study this question by formulating three supervision regimes, which compare monitors trained with safety data from one individual modality ($\mathcal{O}(1)$), every individual modality ($\mathcal{O}(n)$), and every nonempty modality configuration ($\mathcal{O}(2^n)$). Instantiating with text, image, and audio modalities, we evaluate all three regimes on two datasets across five omni-models. The performance differences between successive regimes define the *modality gap* and the *combination gap*. Empirically, the modality gap is generally larger, while much of the combination gap can be recovered with a simple strategy that retains $\mathcal{O}(n)$ safety supervision. To reduce the modality gap, we learn a modality-shared, safety-informed subspace while keeping $\mathcal{O}(1)$ safety supervision. Our method improves mean AUC by 3.2 and 3.9 points on the two datasets, closing 57.1% and 68.4% of the modality gap, respectively. Initialized only once, the learned subspace can be reused across new safety datasets, with strong performance in cross-dataset generalization and continual monitor updating.

1 INTRODUCTION

Large language models (LLMs) have advanced rapidly (Vaswani et al., 2017; Achiam et al., 2023), demonstrating strong capabilities across diverse tasks (Kojima et al., 2022; Li et al., 2022; Wei et al., 2022). However, they remain vulnerable to adversarial attacks that elicit harmful or misleading content (Zou et al., 2023; Liu et al., 2024; Chen et al., 2024). To address this, **safety monitoring** has emerged as a promising solution (Sharma et al., 2025; Oldfield et al., 2026; Cunningham et al., 2026; Kramár et al., 2026; Jiao et al., 2026), which uses additional deployment-time systems to detect harmful user requests. Unlike model-level interventions such as safety fine-tuning and reinforcement learning (Bai et al., 2022; Bianchi et al., 2024), safety monitoring decouples the defense from the underlying model. It can therefore be deployed, updated, and audited independently.

In this work, we focus on *probe-based monitors*, which apply lightweight classifiers to internal hidden representations (Alain & Bengio, 2016; Oldfield et al., 2026; Kramár et al., 2026; Liu et al., 2026). Unlike *LLM-based monitors* (Inan et al., 2023; Zeng et al., 2025), probes directly use the underlying model's representations and avoid an additional LLM inference pass, making them well-suited to low-overhead, always-on monitoring (Kramár et al., 2026). Prior work has explored different probe architectures for LLMs (Oldfield et al., 2026; Kramár et al., 2026). Extending them to omni-models, however, introduces a new combinatorial scaling challenge on the data side.

Omni-models process text, visual, and audio inputs through a shared language-model backbone (Xu et al., 2025a;b; Abouelenin et al., 2025). For safety monitoring, the same harmful request may be expressed through any supported modality or combination, requiring monitors to recognize semantically equivalent requests across different input configurations such as text, text+image, and text+image+audio (Figure 1, left). However, modality-induced representation shifts (Liang et al., 2022; Xiang et al., 2025) mean that a text-trained probe may not generalize well to non-text inputs.

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

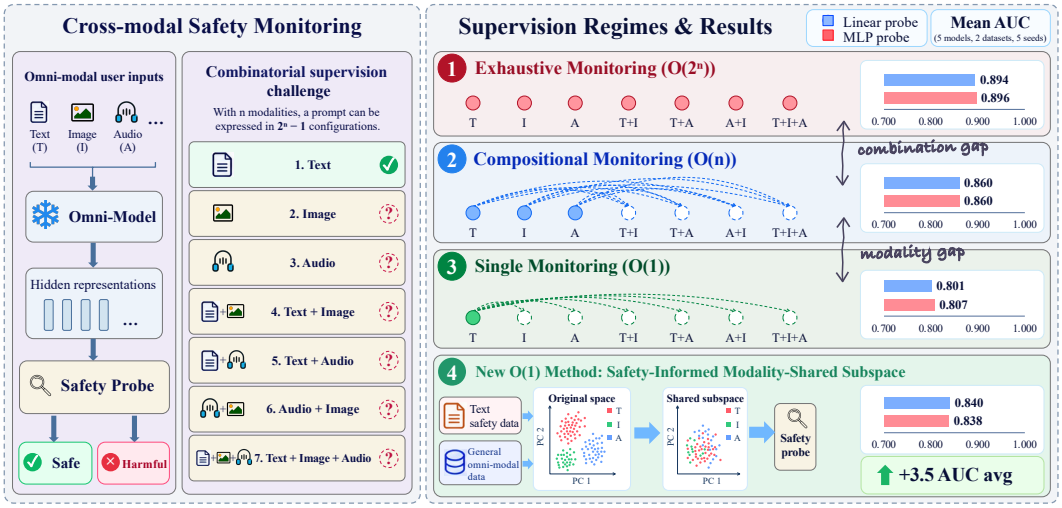


Figure 1: **Overview of cross-modal safety monitoring. Left:** A text-trained monitor may fail to detect harmful requests expressed in other configurations, while training on all configurations is costly. **Right:** We formulate and evaluate three supervision regimes with different training-data coverage, and improve Single Monitoring ($\mathcal{O}(1)$) using a subspace method.

As omni-models incorporate more input modalities[1] (Aghajanyan et al., 2023; Lu et al., 2024), the number of nonempty configurations grows combinatorially: n modalities yield $2^n - 1$ input configurations. Yet safety supervision is unevenly distributed: text safety datasets are comparatively well established, whereas non-text and combined-modality safety data remain scarce and costly to build (Lu et al., 2025). Training on every possible realization of safety prompts, and continually updating such data, therefore becomes increasingly burdensome, motivating our central question:

Can we achieve effective cross-modal safety monitoring in a data-efficient way?

We study this question by formulating three supervision regimes with different training-data coverage (Figure 1, right): *(i)* Exhaustive Monitoring uses safety data from every configuration, requiring $\mathcal{O}(2^n)$ training-data coverage across configurations and serving as a fully supervised reference; *(ii)* Compositional Monitoring uses safety data from the n individual modalities and generalizes to unseen combinations, requiring $\mathcal{O}(n)$ coverage; and *(iii)* Single Monitoring uses safety data from a single modality and generalizes to the remaining configurations, requiring $\mathcal{O}(1)$ coverage.

To evaluate these regimes, we instantiate the framework with text, image, and audio, and extend WildGuardMix (Han et al., 2024) and Aegis 2.0 (Ghosh et al., 2025) to all seven configurations. Across five omni-models, cross-configuration transfer is asymmetric both among singleton modalities and between singleton and combined configurations, highlighting the challenge of cross-modal monitoring. We then compare the three supervision regimes and quantify two gaps: the *modality gap* between $\mathcal{O}(1)$ and $\mathcal{O}(n)$, associated with gains from extending single supervision to all modalities, and the *combination gap* between $\mathcal{O}(n)$ and $\mathcal{O}(2^n)$, associated with additional gains from extending supervision to all modality combinations. Empirically, the modality gap is generally larger.

Based on these findings, we seek to reduce both gaps without increasing supervision complexity. For the combination gap, we find that simply adding the full text+image+audio configuration to singleton training recovers a substantial fraction of the gap, while retaining $\mathcal{O}(n)$ coverage. Reducing the modality gap under the more restrictive $\mathcal{O}(1)$ setting, however, requires safety knowledge to transfer to modalities without direct safety supervision. We therefore learn a *modality-shared, safety-informed subspace* using non-safety-related omni-modal data and text safety data. The former estimates modality-dependent variation, while the latter captures safety-domain semantic variation and benign–harmful separation. We optimize the subspace to suppress modality variation while preserving safety-relevant signals, then project all representations and apply a text-trained probe

[1]Less conventional modalities may also convey safety-relevant content. For example, point clouds may depict unsafe scenes, while skeleton data may represent threatening gestures.

across configurations. This approach improves average AUC by 3.2 points on WildGuardMix-Omni and 3.9 points on Aegis-Omni, recovering 57.1% and 68.4% of the modality gap, respectively. We further show strong performance under both cross-dataset generalization and continual monitor updating, with the subspace initialized only once and directly reused across new safety datasets.

In summary, our contributions are threefold: **(1)** We identify the combinatorial supervision challenge in omni-model safety monitoring and introduce a unified taxonomy of $\mathcal{O}(1)$, $\mathcal{O}(n)$, and $\mathcal{O}(2^n)$ supervision regimes. **(2)** We systematically evaluate these regimes across two datasets and five omni-models, quantify the *modality gap* and *combination gap*, and show that much of the combination gap can be recovered while retaining $\mathcal{O}(n)$ supervision. **(3)** To reduce the modality gap, we learn a modality-shared subspace with text-only safety supervision, improving mean AUC by 3.2 and 3.9 points, which close 57.1% and 68.4% of the modality gap on the two datasets, respectively.

2 Related Work

Omni-Models. Recent work has extended text-centric large language models (LLMs) into omni-models that process text, images, audio, and video (Xu et al., 2025b; Abouelenin et al., 2025). Representative open-source models include Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), and Phi-4-Multimodal (Abouelenin et al., 2025). Omni-models typically use modality-specific encoders (Dosovitskiy et al., 2020; Radford et al., 2023) and projectors to map non-text features into a shared LLM backbone, whose hidden representations provide a natural interface for safety monitoring. Alignment training encourages representations of semantically equivalent inputs to align across modalities, although the degree of alignment varies across models (Chen et al., 2026).

Safety Monitoring. LLMs are post-trained to produce safe outputs (Bai et al., 2022; Dai et al., 2024; Bianchi et al., 2024; Yuan et al., 2025), yet remain vulnerable to adversarial attacks (Zou et al., 2023; Liu et al., 2024; Hughes et al., 2025). Many systems therefore employ an additional LLM to screen prompts and outputs (Inan et al., 2023; Han et al., 2024; Zeng et al., 2025). Although effective, it introduces substantial computational overhead for always-on monitoring and edge deployment (Oldfield et al., 2026; Team et al., 2026). Probe-based monitors offer a lightweight alternative by classifying hidden representations (Park et al., 2024; Sharma et al., 2025; Oldfield et al., 2026; Liu et al., 2026; Jiao et al., 2026). Prior work has largely focused on probe architectures, exploring linear probes (Alain & Bengio, 2016), MLPs (Teerapittayanon et al., 2016), bilinear probes (Hewitt & Liang, 2019), attention probes (Kramár et al., 2026), and Truncated Polynomial Classifiers (Oldfield et al., 2026). The *data perspective*, particularly how much safety supervision is required for generalization across modality configurations, remains underexplored. Recent work has begun to study cross-modal safety transfer. SafeSteer (Zeng et al., 2026) transfers text safety alignment to images, while Multiguard (Verma et al., 2025) identifies aligned representations for image and audio safety. Omni-SafetyBench (Pan et al., 2025a) and AdvBench-Omni (Wang et al., 2026) evaluate safety across individual and combined modalities. However, these works do not systematically study how supervision coverage across modality configurations scales for cross-modal monitoring. We address this question through a supervision hierarchy covering a single modality, all singleton modalities, and all modality configurations, scaling as $\mathcal{O}(1)$, $\mathcal{O}(n)$, and $\mathcal{O}(2^n)$, respectively, and investigate how to improve performance under the more data-efficient $\mathcal{O}(1)$ and $\mathcal{O}(n)$ regimes.

Safety Subspace Learning. Prior work suggests that safety-related behavior is represented in low-dimensional activation geometry, ranging from refusal directions to multidimensional safety subspaces used for analysis or intervention (Arditi et al., 2024; Pan et al., 2025b; Fu et al., 2026). Most closely related to our $\mathcal{O}(1)$ method, Fu et al. (2026) uses text and image safety data to construct *modality-specific* subspaces for activation intervention. Our method instead combines non-safety-related omni-modal data with text-only safety data to learn a *modality-shared*, safety-informed subspace for effective cross-modal safety monitoring.

3 Cross-Modal Safety Monitoring from a Data Perspective

3.1 Problem Formulation

Let $\mathcal{M} = \{m_1, \ldots, m_n\}$ denote the n input modalities supported by an omni-model. Each non-empty subset $c \subseteq \mathcal{M}$ defines a *modality configuration*, giving the configuration space $\mathcal{C} = 2^{\mathcal{M}} \setminus \{\emptyset\}$

with $|\mathcal{C}| = 2^n - 1$. For each configuration c, let $\mathcal{D}^c = \{(\mathbf{x}_i^c, y_i)\}_{i=1}^{N^c}$ denote its safety dataset of size N^c, where $\mathbf{x}_i^c$ is a prompt expressed through the modalities in c, and $y_i \in \{0, 1\}$ indicates whether it is benign (0) or harmful (1). For each input, we extract a pooled hidden representation $\mathbf{h}_i^c \in \mathbb{R}^d$ from a layer of the omni-model's LLM backbone. A safety probe $f : \mathbb{R}^d \to \mathbb{R}$ maps this representation to a harmfulness score. Let $\mathcal{T} \subseteq \mathcal{C}$ denote the available configurations used for safety training. Our objective is to learn monitors from $\{\mathcal{D}^c : c \in \mathcal{T}\}$ that perform well across all configurations in $\mathcal{C}$.

Supervision coverage. We distinguish two dimensions of safety supervision: *modality coverage*, $\kappa_{\text{mod}}(\mathcal{T}) = \left|\bigcup_{c \in \mathcal{T}} c\right|$, counts the distinct modalities represented in safety training; *configuration coverage*, $\kappa_{\text{cfg}}(\mathcal{T}) = |\mathcal{T}|$, counts the distinct supervised configurations. We report coverage pairs as $(\kappa_{\text{mod}}, \kappa_{\text{cfg}})$. Configuration coverage measures the number of configuration-specific safety datasets required, whereas modality coverage reflects the breadth of modalities involved in constructing them. For example, training only on the full joint configuration, $\mathcal{T} = \{\mathcal{M}\}$, has coverage $(n, 1)$: it requires one safety dataset containing inputs that span all n modalities (e.g., text+image+audio when $n = 3$). We primarily focus on configuration coverage to study how the number of required datasets scales with n while keeping modality coverage constrained.

Supervision regimes. We focus on three supervision regimes. *Exhaustive Monitoring* uses $\mathcal{T}_{\text{full}} = \mathcal{C}$, with coverage $(n, 2^n - 1)$, providing safety data for every configuration. *Compositional Monitoring* uses $\mathcal{T}_{\text{comp}} = \{\{m_1\}, \ldots, \{m_n\}\}$, with coverage (n, n), and generalizes from singleton configurations to combined-modality configurations. *Single Monitoring* uses $\mathcal{T}_{\text{single}} = \{c_0\}$ for one singleton configuration c_0, with coverage $(1, 1)$, and transfers its monitor to all remaining configurations. Their configuration coverage scales as $\mathcal{O}(2^n)$, $\mathcal{O}(n)$, and $\mathcal{O}(1)$, respectively; throughout the paper, we use these orders as shorthand for the corresponding regimes.

Empirical Setup. We instantiate our framework with $n = 3$ modalities (text, image, and audio), with $\mathcal{M} = \{m_{\text{t}}, m_{\text{i}}, m_{\text{a}}\}$, yielding seven non-empty configurations. Then the supervision regimes become $\mathcal{T}_{\text{full}} = \{\{m_{\text{t}}\}, \{m_{\text{i}}\}, \{m_{\text{a}}\}, \{m_{\text{t}}, m_{\text{i}}\}, \{m_{\text{t}}, m_{\text{a}}\}, \{m_{\text{i}}, m_{\text{a}}\}, \{m_{\text{t}}, m_{\text{i}}, m_{\text{a}}\}\}$, $\mathcal{T}_{\text{comp}} = \{\{m_{\text{t}}\}, \{m_{\text{i}}\}, \{m_{\text{a}}\}\}$, and $\mathcal{T}_{\text{single}} = \{\{m_{\text{t}}\}\}$. For *Single Monitoring*, we choose text as the source modality for two reasons: text safety datasets are substantially more developed than their non-text counterparts, and omni-models are typically text-centric, mapping non-text features into the representation space of an LLM backbone. Text therefore provides a natural source of $\mathcal{O}(1)$ safety supervision and a common representational anchor for cross-configuration transfer.

3.2 Evaluation of Cross-Configuration Generalization

To evaluate these supervision regimes, we construct WildGuardMix-Omni (WGM-Omni) and Aegis-Omni from the two text safety datasets WildGuardMix (Han et al., 2024) and Aegis AI Content Safety Dataset 2.0 (Ghosh et al., 2025), respectively. The constructed omni datasets cover all seven modality configurations considered in our study. We train a monitor on each configuration and evaluate it across all configurations, allowing us to characterize cross-configuration generalization and obtain configuration-specific monitors for subsequent ensembling and comparison.

Dataset Construction. For each source dataset, we randomly sample 10,000 prompts from the training split and use the full test split. We then convert each prompt into all seven configurations while preserving its original semantics. For modality combinations, the semantic content of the original prompt is distributed across the constituent modalities. For instance, a prompt "How to make a bomb" is transformed into the text "How to make the object in the image" paired with an image rendering the word "bomb". Construction details are provided in Appendix A.1. This deliberately simple and controlled construction isolates the question studied in this work: *whether a monitor can correctly classify the same prompt when its semantic content is re-expressed across different modality configurations*. By minimizing changes to the underlying content, performance differences are less confounded by semantic variation across configurations. Notably, our results show that generalization degrades even under these simple transformations, highlighting the challenge of cross-configuration transfer even before considering complex attack variations. Our construction therefore focuses on this basic transfer setting rather than the full diversity of omni-modal interactions.

Experimental Setup. We evaluate five open-source omni-models spanning different architectures and model sizes: `Qwen2.5-Omni-3B` and `Qwen2.5-Omni-7B` (Xu et al., 2025a), `Qwen3-Omni-30B` (Xu et al., 2025b), `Phi4-mm-5.6B` (Abouelenin et al., 2025), and `InteractiveOmni-8B` (Tong et al., 2025). All five models support text, image, audio, and

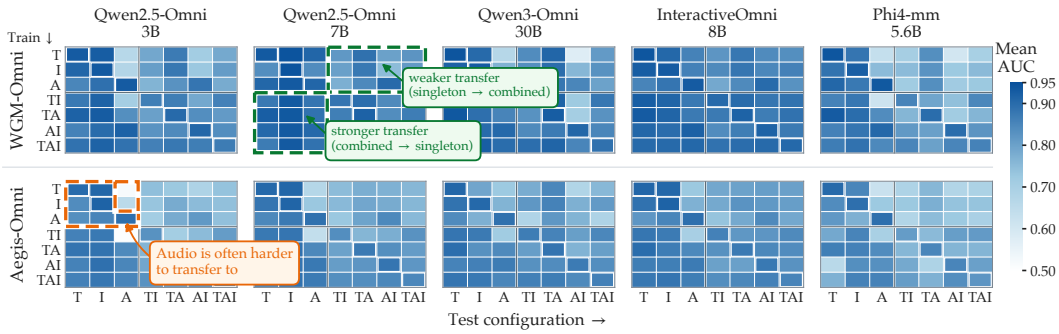


Figure 2: **Cross-configuration generalization results on WGM-Omni and Aegis-Omni.** We show the results of the linear probe here. Full results are visualized in Figure 5.

their combinations as input. Following prior work (Liu et al., 2026; Oldfield et al., 2026), we extract pooled residual-stream activations from the LLM backbone, focusing on later layers due to their encoding of high-level semantic information (Park et al., 2024). We evaluate both linear and MLP (two-layer) probes, representing monitor architectures with different levels of expressiveness. Experiments are conducted with five random seeds. We report AUROC (AUC) as our primary metric, which does not require a fixed decision threshold. Complementary metrics and experimental details are provided in Appendices B.2 and A.2.1, respectively.

Results and Analysis. Figure 2 visualizes cross-configuration generalization of linear probes across five omni-models and two datasets (with full numerical results reported in Tables 12 and 13). We observe three main findings. (1) *Transfer across singleton configurations is asymmetric.* Among text, image, and audio, audio is generally the most difficult target to generalize to, highlighting the challenge of monitor generalization across modalities. (2) *Transfer between singleton and combined configurations is also asymmetric.* Monitors trained on combined configurations generally transfer better to their constituent singleton configurations than singleton-trained monitors transfer in the reverse direction. For example, a monitor trained on text+image+audio typically generalizes better to text, image, and audio individually than monitors trained on these singleton modalities generalize to text+image+audio. (3) *Increasing probe expressiveness does not consistently improve generalization.* As shown in Figure 6, MLP probes show no consistent advantage over linear probes across the seven test configurations, and often perform worse. We further study increasingly expressive MLP probes in Figure 7, where greater probe capacity yields little or no additional improvement in generalization. These results suggest that probe expressiveness is not the primary bottleneck, motivating our focus on the data perspective. Additional analysis is provided in Appendix B.1.

3.3 Implementation of Three Supervision Regimes

We implement the three supervision regimes defined in Section 3.1 using the monitors trained in Section 3.2. Each regime specifies the safety training-data coverage across configurations, while allowing different choices for the number of monitors and how their predictions are combined at inference. For *Single Monitoring* ($\mathcal{O}(1)$), we apply the text-trained monitor to all seven test configurations. For *Compositional Monitoring* ($\mathcal{O}(n)$), we evaluate four ensemble rules over the text-, image-, and audio-trained monitors: average probability, compositional average probability, maximum probability, and majority voting. Compositional average probability uses only monitors corresponding to the constituent modalities of the test configuration. We additionally consider a single monitor trained on the pooled text, image, and audio datasets, denoted *One Monitor Pooled Data*. For *Exhaustive Monitoring* ($\mathcal{O}(2^n)$), we evaluate average probability, maximum probability, and majority voting over all configuration-specific monitors, as well as a single monitor trained on the union of all seven datasets. We additionally consider *Configuration-Matched* inference, which directly applies the monitor trained on the same modality configuration as the test input. For each supervision regime, we report its overall best-performing variant in the main comparison and show results for all variants in Tables 18–22.

3.4 Analysis of Modality Gap and Combination Gap

Figure 3 compares the three supervision regimes across five omni-models on WGM-Omni and Aegis-Omni, with numerical results reported in Tables 29 and 32. The supervision hierarchy al-

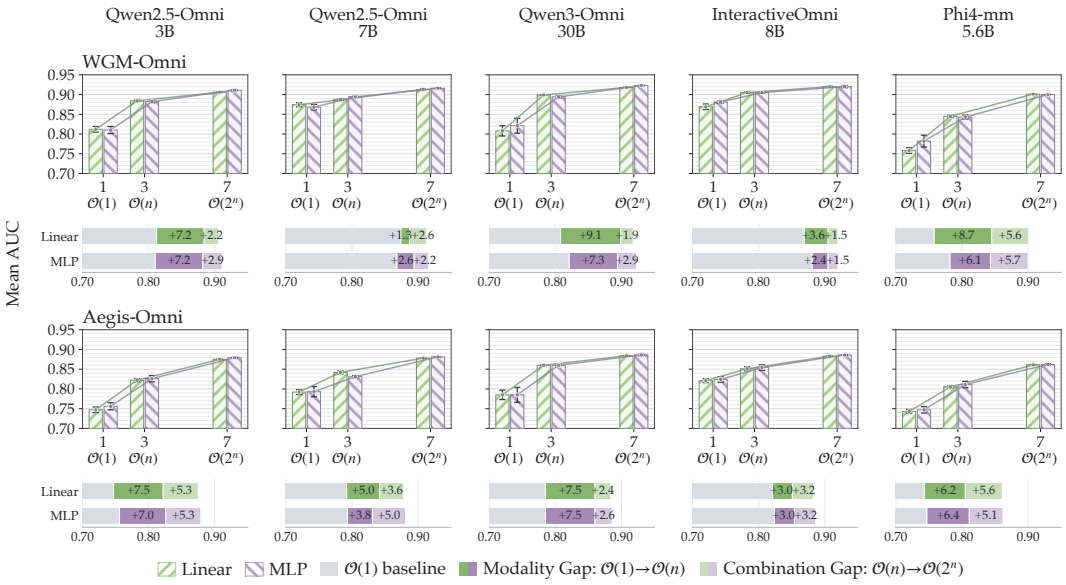


Figure 3: **Results of three supervision regimes on WGM-Omni and Aegis-Omni.** Beneath each panel, stacked bars decompose mean AUC into the $\mathcal{O}(1)$ baseline and the incremental gains to $\mathcal{O}(n)$ and $\mathcal{O}(2^n)$ for Linear and MLP probes. Gain labels are in AUC points ($100 \times \Delta\text{AUC}$).

lows us to quantify two distinct generalization gaps. We define the performance difference between $\mathcal{O}(1)$ and $\mathcal{O}(n)$ as the **modality gap**, which measures the empirical gain associated with extending safety supervision from text alone to all singleton modalities. Similarly, we define the difference between $\mathcal{O}(n)$ and $\mathcal{O}(2^n)$ as the **combination gap**, which measures the additional gain associated with extending supervision from singleton modalities to all modality configurations.

Empirically, the modality gap is generally larger, exceeding the combination gap in most cases, particularly for `Qwen2.5-Omni-3B`, `Qwen3-Omni-30B`, and `Phi4-mm` (Figure 3). This suggests that extending safety supervision to unseen modalities generally provides a larger performance gain than additionally supervising unseen combinations of already covered modalities. Moreover, as we show next, much of the combination gap can be recovered without exhaustive $\mathcal{O}(2^n)$ supervision.

Reducing the Combination Gap. We next ask whether the combination gap can be reduced while keeping $\mathcal{O}(n)$ supervision. To this end, we augment the singleton text, image, and audio training data with one joint configuration, text+audio+image (TAI). Using the notation of Section 3.1, the augmented supervision set is $\mathcal{T}_{\text{aug}} = \{\{m_{\text{t}}\}, \{m_{\text{i}}\}, \{m_{\text{a}}\}\} \cup \{\{m_{\text{t}}, m_{\text{i}}, m_{\text{a}}\}\}$, for which $\kappa_{\text{mod}}(\mathcal{T}_{\text{aug}}) = n$ and $\kappa_{\text{cfg}}(\mathcal{T}_{\text{aug}}) = n + 1 = \mathcal{O}(n)$. Thus, modality coverage remains unchanged, and configuration coverage remains $\mathcal{O}(n)$. For clarity in our comparisons, we denote this augmented setting as $\mathcal{O}(n+1)$. Despite this modest addition, the resulting monitor recovers a substantial fraction of the combination gap (see Tables 1 and 2). On WGM-Omni, it improves mean AUC by 2.3 points over the original $\mathcal{O}(n)$ baseline, closing 79.0% of the combination gap. On Aegis-Omni, the improvement is 2.7 AUC points, closing 64.9% of the gap.

Reducing the modality gap, however, is more challenging. Unlike the combination setting, where every modality already receives direct safety supervision, the $\mathcal{O}(1)$ setting requires safety knowledge learned from text to generalize to modalities with no direct safety supervision. This setting is also practically important because existing safety datasets are highly imbalanced across modalities, with substantially more supervision available for text than for non-text modalities. We therefore next investigate how to reduce the modality gap while retaining $\mathcal{O}(1)$ safety supervision.

4 Improving $\mathcal{O}(1)$ Generalization

Our analysis identifies the modality gap as the main challenge for text-trained monitors. To improve transfer while retaining $\mathcal{O}(1)$ safety supervision, we seek a *modality-shared*, *safety-informed* subspace with three desired properties: representations of the same prompt have little modality-specific

variation, representations of different prompts preserve semantic variation, and benign and harmful text prompts remain well separated. We capture these properties using three covariance matrices: $\mathbf{S}_{\text{modality}}$, $\mathbf{S}_{\text{semantic}}$, and $\mathbf{S}_{\text{safety}}$. The key insight is that $\mathbf{S}_{\text{modality}}$ and $\mathbf{S}_{\text{semantic}}$ can be estimated from general, non-safety-related data, allowing safety supervision to remain $\mathcal{O}(1)$. In practice, though, we find that estimating $\mathbf{S}_{\text{semantic}}$ from text safety data yields better performance (Table 4).

4.1 Modality-shared Safety-Informed Subspace

Modality Variation. Following the notation in Section 3.1, let $\mathcal{C}_1 = \{\{m_1\}, \ldots, \{m_n\}\}$ denote the singleton modalities. We use N general prompts, each expressed in every singleton configuration $c \in \mathcal{C}_1$. Recall that $\mathbf{h}_i^c \in \mathbb{R}^d$ denotes the pooled representation of prompt i in configuration c. We first center each singleton modality by its empirical mean $\boldsymbol{\mu}_c = \frac{1}{N}\sum_{i=1}^{N} \mathbf{h}_i^c$, obtaining $\tilde{\mathbf{h}}_i^c = \mathbf{h}_i^c - \boldsymbol{\mu}_c$. For each prompt, we compute its mean representation across modalities, $\mathbf{z}_i = \frac{1}{n}\sum_{c\in\mathcal{C}_1} \tilde{\mathbf{h}}_i^c$, and the modality-specific residual $\mathbf{r}_i^c = \tilde{\mathbf{h}}_i^c - \mathbf{z}_i$. The modality covariance is then

$$\mathbf{S}_{\text{modality}} = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{c\in\mathcal{C}_1} \mathbf{r}_i^c (\mathbf{r}_i^c)^\top.$$

Safety Conditioning. Let $\mathcal{J}$ index labeled text safety examples with representations $\mathbf{h}_j^{c_\text{t}}$, where $c_\text{t} = \{m_\text{t}\}$, and labels $y_j \in \{0, 1\}$ indicating benign and harmful prompts, respectively. We define

$$\mathbf{S}_{\text{semantic}} = \text{Cov}_{j\in\mathcal{J}}\left(\mathbf{h}_j^{c_\text{t}}\right), \qquad \mathbf{S}_{\text{safety}} = \pi_0\pi_1(\boldsymbol{\nu}_1 - \boldsymbol{\nu}_0)(\boldsymbol{\nu}_1 - \boldsymbol{\nu}_0)^\top.$$

Here, Cov denotes the empirical covariance, while $\boldsymbol{\nu}_y$ and π_y, for $y \in \{0, 1\}$, denote the representation centroid and empirical frequency of class y, respectively. Under equal class weighting, $\pi_0 = \pi_1 = 1/2$. The semantic covariance captures variation among safety-domain prompts, while the rank-one safety covariance captures the direction separating the benign and harmful centroids.

Subspace Estimation. To balance their scales, we normalize each covariance by its average marginal variance $\widehat{\mathbf{S}} = \frac{\mathbf{S}}{\text{tr}(\mathbf{S})/d+\epsilon}$, where $\epsilon > 0$ ensures numerical stability. For a direction $\mathbf{v} \in \mathbb{R}^d$, we define the safety-informed shared-subspace criterion using the generalized Rayleigh quotient $\frac{\mathbf{v}^\top(\widehat{\mathbf{S}}_{\text{semantic}}+\alpha\widehat{\mathbf{S}}_{\text{safety}})\mathbf{v}}{\mathbf{v}^\top(\widehat{\mathbf{S}}_{\text{modality}}+\rho\mathbf{I}_d)\mathbf{v}}$. The numerator favors directions that preserve variation within the safety domain and separation between benign and harmful prompts, while the denominator penalizes modality-specific variation. Here, $\alpha \geq 0$ controls the strength of safety conditioning, and the ridge coefficient $\rho > 0$ ensures that the denominator matrix is positive definite. Maximizing this criterion yields the generalized eigenvalue problem

$$\left(\widehat{\mathbf{S}}_{\text{semantic}} + \alpha\widehat{\mathbf{S}}_{\text{safety}}\right)\mathbf{v}_j = \lambda_j\left(\widehat{\mathbf{S}}_{\text{modality}} + \rho\mathbf{I}_d\right)\mathbf{v}_j.$$

We construct $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_k] \in \mathbb{R}^{d\times k}$ from the generalized eigenvectors associated with the k largest eigenvalues, normalized such that $\mathbf{V}^\top\left(\widehat{\mathbf{S}}_{\text{modality}} + \rho\mathbf{I}_d\right)\mathbf{V} = \mathbf{I}_k$. The columns of $\mathbf{V}$ span the resulting k-dimensional subspace. We then train a safety monitor on the projected text features $\mathbf{p}^{c_\text{t}} = \mathbf{V}^\top(\mathbf{h}^{c_\text{t}} - \boldsymbol{\mu}_{c_\text{t}})$. At inference, for every configuration $c \in \mathcal{C}$, we project $\mathbf{p}^c = \mathbf{V}^\top(\mathbf{h}^c - \boldsymbol{\mu}_{c_\text{t}})$ and apply the same text-trained monitor. We use the fixed text mean $\boldsymbol{\mu}_{c_\text{t}}$ from the general prompts, as test-configuration statistics are unavailable in the generalization setting.

4.2 Experimental Setup

We randomly sample 5,000 prompts from NoRobots (Rajani et al., 2023), a general, non-safety-related dataset, and convert them into image and audio inputs. While paired data are widely used in omni-model training (Xu et al., 2025a; Abouelenin et al., 2025), complete training corpora are often unavailable. We therefore construct these variants ourselves. We further evaluate sensitivity to the choice of general dataset in Appendix C.3.3, showing that the method is not specific to NoRobots. NoRobots representations are used to estimate $\mathbf{S}_{\text{modality}}$, while text safety representations are used to estimate $\mathbf{S}_{\text{semantic}}$ and $\mathbf{S}_{\text{safety}}$. For hyperparameter selection, we reserve subsets of both data sources for validation and grid search over (α, ρ, k). Further details are provided in Appendix A.2.2.

Table 1: **Mean AUC across supervision regimes on WGM-Omni.** Green cells mark our $\mathcal{O}(1)$ method; the green ↑ value is its gain in AUC points over the corresponding original $\mathcal{O}(1)$ result. Entries report mean ± sample standard deviation across five seeds, rounded to three decimals.

Monitor	Method	Qwen2.5-Omni-3B	Qwen2.5-Omni-7B	Qwen3-Omni-30B	Phi4-mm	Interactive-Omni-8B
Linear	$\mathcal{O}(1)$	0.812 ± 0.007	0.874 ± 0.005	0.808 ± 0.013	0.758 ± 0.007	0.869 ± 0.007
	$\mathcal{O}(1)$ - **Ours**	**0.852** ± 0.002 ↑4.0	**0.874** ± 0.003 ↑0.0	**0.870** ± 0.002 ↑6.2	**0.798** ± 0.009 ↑4.0	**0.879** ± 0.006 ↑1.0
	$\mathcal{O}(n)$	0.884 ± 0.002	0.887 ± 0.002	0.899 ± 0.001	0.845 ± 0.003	0.905 ± 0.002
	$\mathcal{O}(n+1)$	0.900 ± 0.001	0.909 ± 0.002	0.912 ± 0.001	0.883 ± 0.001	0.912 ± 0.002
	$\mathcal{O}(2^n)$	0.906 ± 0.001	0.913 ± 0.001	0.918 ± 0.001	0.901 ± 0.000	0.920 ± 0.001
MLP	$\mathcal{O}(1)$	0.810 ± 0.009	0.868 ± 0.007	0.821 ± 0.019	0.782 ± 0.015	0.881 ± 0.003
	$\mathcal{O}(1)$ - **Ours**	**0.869** ± 0.001 ↑5.9	**0.879** ± 0.005 ↑1.1	**0.869** ± 0.002 ↑4.8	**0.823** ± 0.006 ↑4.1	**0.887** ± 0.005 ↑0.6
	$\mathcal{O}(n)$	0.882 ± 0.003	0.894 ± 0.002	0.894 ± 0.003	0.843 ± 0.005	0.905 ± 0.002
	$\mathcal{O}(n+1)$	0.910 ± 0.001	0.917 ± 0.002	0.917 ± 0.001	0.891 ± 0.002	0.916 ± 0.003
	$\mathcal{O}(2^n)$	0.911 ± 0.001	0.916 ± 0.001	0.923 ± 0.001	0.900 ± 0.000	0.920 ± 0.003

Table 2: **Mean AUC across supervision regimes on Aegis-Omni.** Green cells mark our $\mathcal{O}(1)$ method; the green ↑ value is its gain in AUC points over the corresponding original $\mathcal{O}(1)$ result. Entries report mean ± sample standard deviation across five seeds, rounded to three decimals.

Monitor	Method	Qwen2.5-Omni-3B	Qwen2.5-Omni-7B	Qwen3-Omni-30B	Phi4-mm	Interactive-Omni-8B
Linear	$\mathcal{O}(1)$	0.747 ± 0.007	0.792 ± 0.006	0.785 ± 0.012	0.743 ± 0.005	0.821 ± 0.005
	$\mathcal{O}(1)$ - **Ours**	**0.810** ± 0.002 ↑6.3	**0.825** ± 0.001 ↑3.3	**0.831** ± 0.001 ↑4.6	**0.812** ± 0.003 ↑6.9	**0.852** ± 0.001 ↑3.1
	$\mathcal{O}(n)$	0.822 ± 0.004	0.842 ± 0.004	0.860 ± 0.001	0.805 ± 0.003	0.851 ± 0.005
	$\mathcal{O}(n+1)$	0.859 ± 0.001	0.866 ± 0.002	0.868 ± 0.001	0.846 ± 0.002	0.870 ± 0.005
	$\mathcal{O}(2^n)$	0.875 ± 0.001	0.878 ± 0.001	0.884 ± 0.001	0.861 ± 0.001	0.883 ± 0.001
MLP	$\mathcal{O}(1)$	0.756 ± 0.009	0.793 ± 0.013	0.785 ± 0.019	0.747 ± 0.008	0.824 ± 0.007
	$\mathcal{O}(1)$ - **Ours**	**0.809** ± 0.006 ↑5.3	**0.816** ± 0.003 ↑2.3	**0.813** ± 0.006 ↑2.8	**0.771** ± 0.006 ↑2.4	**0.843** ± 0.001 ↑1.9
	$\mathcal{O}(n)$	0.826 ± 0.008	0.831 ± 0.003	0.860 ± 0.001	0.811 ± 0.008	0.854 ± 0.007
	$\mathcal{O}(n+1)$	0.863 ± 0.006	0.865 ± 0.003	0.873 ± 0.002	0.849 ± 0.001	0.871 ± 0.008
	$\mathcal{O}(2^n)$	0.879 ± 0.001	0.881 ± 0.001	0.886 ± 0.002	0.862 ± 0.002	0.886 ± 0.001

4.3 Results and Analysis

Tables 1 and 2 report five-seed results on WGM-Omni and Aegis-Omni, respectively, with full results in Tables 29 and 32 and further analysis provided in Appendix C.3.

(1) Our method provides broad gains under $\mathcal{O}(1)$ safety supervision. Across five models and two probe architectures, our method closes a substantial fraction of the modality gap (Figure 4). Specifically, it improves mean AUC by 3.2 points on WGM-Omni and 3.9 points on Aegis-Omni, reducing the mean modality gap by 57.1% and 68.4%, respectively. The only exception is `Qwen2.5-Omni-7B` with a linear probe on WGM-Omni, where performance remains unchanged, and the original modality gap is already small at 1.3 AUC points. Additional sensitivity analyses are provided in Appendix C.3.3.

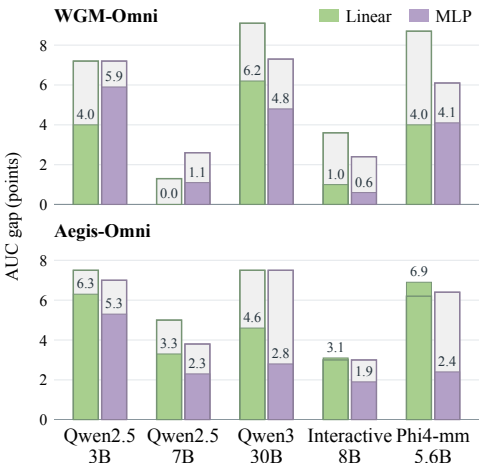


Figure 4: **Recovery of the modality gap in AUC points.** Colored bars show actual values; gray bars denote the original gap.

(2) The learned subspace improves audio generalization while largely preserving text performance, with benefits extending to modality combinations. Tables 29 and 32 show that our method improves audio AUC in most settings, with particularly large gains when the original transfer to audio is weak (`Qwen2.5-Omni-3B`, `Qwen3-Omni-30B`, and `Phi4-mm`). These gains come with text-AUC reductions of at most 1.6 points across both datasets, indicating only a modest trade-off between source-modality performance and cross-configuration generalization. Improvements also extend to combined configurations, particularly on Aegis-Omni, suggesting that a subspace learned from singleton modalities can also improve generalization to modality combinations.

(3) Our method achieves strong cross-dataset generalization while requiring only one-time subspace learning for continual monitor updating. We first evaluate cross-dataset generalization of the $\mathcal{O}(1)$

Table 3: **Cross-dataset generalization and continual updating from WGM-Omni to Aegis-Omni (AUC).** The first two rows show generalization results; the last two show updating results.

Monitor	Method	Qwen2.5-Omni-3B	Qwen2.5-Omni-7B	Qwen3-Omni-30B	Phi4-mm	Interactive-Omni-8B
Linear	$\mathcal{O}(1)$	0.637 ± 0.005	0.737 ± 0.005	0.616 ± 0.006	0.566 ± 0.009	0.689 ± 0.008
	$\mathcal{O}(1)$ **- Ours**	**0.700** ± 0.003 ↑6.3	**0.747** ± 0.004 ↑1.0	**0.719** ± 0.004 ↑10.3	**0.639** ± 0.010 ↑7.3	**0.766** ± 0.011 ↑7.7
	$\mathcal{O}(1)$ - Updated	0.696 ± 0.005	0.778 ± 0.004	0.690 ± 0.006	0.627 ± 0.009	0.748 ± 0.008
	$\mathcal{O}(1)$ **- Ours - Updated**	**0.777** ± 0.002 ↑8.1	**0.796** ± 0.005 ↑1.8	**0.777** ± 0.010 ↑8.7	**0.722** ± 0.009 ↑9.5	**0.812** ± 0.002 ↑6.4
MLP	$\mathcal{O}(1)$	0.637 ± 0.011	0.728 ± 0.011	0.634 ± 0.009	0.605 ± 0.025	0.705 ± 0.011
	$\mathcal{O}(1)$ **- Ours**	**0.705** ± 0.011 ↑6.8	**0.765** ± 0.006 ↑3.7	**0.729** ± 0.006 ↑9.5	**0.669** ± 0.007 ↑6.4	**0.783** ± 0.017 ↑7.8
	$\mathcal{O}(1)$ - Updated	0.721 ± 0.014	0.789 ± 0.009	0.697 ± 0.006	0.650 ± 0.008	0.767 ± 0.006
	$\mathcal{O}(1)$ **- Ours - Updated**	**0.762** ± 0.008 ↑4.1	**0.800** ± 0.002 ↑1.1	**0.770** ± 0.005 ↑7.3	**0.727** ± 0.005 ↑7.7	**0.817** ± 0.006 ↑5.0

methods from WGM-Omni to Aegis-Omni. For our method, we directly reuse the subspace learned from WGM-Omni to project Aegis-Omni representations and apply the corresponding WGM-Omni text-trained monitors. As shown in Table 3, our method achieves higher average AUC across all models and probe architectures, demonstrating strong generalization to a new safety dataset without re-estimating the subspace. We further consider a practical setting where monitors are continually updated with Aegis-Omni text data. Each monitor is updated using 5,000 balanced Aegis-Omni text examples together with 5,000 balanced WGM-Omni text examples for replay, while our previously learned subspace remains fixed. Table 3 shows that our method retains strong performance after updating. These results highlight a practical advantage: the subspace can be learned once using general omni-modal data and reused thereafter, while subsequent adaptation to new safety distributions requires only text safety supervision. Full results are reported in Table 25.

Ablations. We ablate the three covariance components, $\mathbf{S}_{\text{safety}}$, $\mathbf{S}_{\text{modality}}$, and $\mathbf{S}_{\text{semantic}}$, by removing each in turn. We also estimate $\mathbf{S}_{\text{semantic}}$ from general data rather than text safety data. All variants retain text-only safety supervision and follow the same hyperparameter selection procedure. As shown in Table 4, our method achieves the highest AUC across all models and monitors. Full per-configuration results are reported in Table 23.

Table 4: **Ablations of safety, semantic, and modality covariance.** Average AUC across seven test configurations for our method and four covariance ablations on WGM-Omni.

Monitor	Method	Qwen2.5-Omni-3B	Qwen2.5-Omni-7B	Qwen3-Omni-30B	Phi4-mm	Interactive-Omni-8B
Linear	w/o $\mathbf{S}_{\text{safety}}$	0.830 ± 0.001	0.838 ± 0.001	0.825 ± 0.001	0.733 ± 0.001	0.834 ± 0.001
	w/o $\mathbf{S}_{\text{modality}}$	0.797 ± 0.009	0.852 ± 0.002	0.806 ± 0.009	0.748 ± 0.006	0.865 ± 0.001
	w/o $\mathbf{S}_{\text{semantic}}$	0.801 ± 0.008	0.818 ± 0.017	0.809 ± 0.004	0.785 ± 0.006	0.815 ± 0.011
	General $\mathbf{S}_{\text{semantic}}$	0.817 ± 0.002	0.845 ± 0.001	0.837 ± 0.002	0.795 ± 0.001	0.855 ± 0.001
	Ours	**0.852** ± 0.002	**0.874** ± 0.003	**0.870** ± 0.002	**0.798** ± 0.009	**0.879** ± 0.006
MLP	w/o $\mathbf{S}_{\text{safety}}$	0.855 ± 0.003	0.870 ± 0.003	0.853 ± 0.006	0.792 ± 0.005	0.877 ± 0.001
	w/o $\mathbf{S}_{\text{modality}}$	0.806 ± 0.008	0.854 ± 0.005	0.812 ± 0.012	0.744 ± 0.013	0.863 ± 0.009
	w/o $\mathbf{S}_{\text{semantic}}$	0.798 ± 0.016	0.819 ± 0.010	0.755 ± 0.014	0.771 ± 0.009	0.824 ± 0.014
	General $\mathbf{S}_{\text{semantic}}$	0.846 ± 0.004	0.858 ± 0.002	0.836 ± 0.008	0.802 ± 0.003	0.874 ± 0.002
	Ours	**0.869** ± 0.001	**0.879** ± 0.005	**0.869** ± 0.002	**0.823** ± 0.006	**0.887** ± 0.005

5 Conclusion

Our work provides initial insights into cross-modal safety monitoring from a data perspective, asking how to achieve effective monitoring with data-efficient supervision. We formulate three supervision regimes and quantify the modality and combination gaps, with the modality gap emerging as the larger challenge. To reduce the modality gap and improve $\mathcal{O}(1)$ monitoring, we introduce a modality-shared, safety-informed subspace that retains $\mathcal{O}(1)$ safety supervision while improving mean AUC by 3.2 points on WildGuardMix-Omni and 3.9 points on Aegis-Omni, respectively.

Limitations. We discuss several limitations. First, our evaluation covers only text, image, and audio; extending to video and other modalities remains future work. Second, we focus on cross-modal transfer of semantically equivalent safety prompts under controlled modality transformations. Our results show that generalization already degrades in this basic setting, but we do not evaluate more complex transformations where the original semantics are expressed more indirectly, such as through generated images or dedicated jailbreaking for modality combinations. Third, we focus on safety supervision coverage rather than monitor architecture. Thus, we use linear and MLP probes as representative architectures rather than exhaustively exploring other monitor designs.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, provide feedback on research methodology or experiments, implement methods, and clean and reformat datasets. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, assist with translation, formulate mathematical claims, provide critical ingredients for proving mathematical claims, support qualitative and thematic data analysis, propose or refine hypotheses, interpret results, or assist in the writing of proofs. Additionally, we used generative AI tools to create or modify scientific figures or images and to edit the research paper to improve readability. We have reviewed all AI-assisted work. LLM-written code has been reviewed and verified by the authors. Figures have also been checked and refined. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to improve cross-modal safety monitoring in a data-efficient manner. By comparing three supervision regimes with different levels of modality-configuration coverage and improving performance under the $\mathcal{O}(1)$ and $\mathcal{O}(n)$ settings, we provide insights into effectively monitoring omni-modal inputs using lightweight probes. We hope these findings can reduce the amount of modality-specific safety supervision required for monitoring and contribute to the development of practical guardrails for more trustworthy omni-models. We acknowledge that the WGM-Omni and Aegis-Omni datasets constructed in this work have potential dual-use risks. Although these datasets are intended for defensive research, such as evaluating and improving omni-modal safety mechanisms, their harmful examples could potentially be misused to circumvent existing safety systems. Our dataset construction follows established practices in prior omni-modal safety research, and we will take measures to support responsible release and use. In particular, access to the datasets will require users to agree to responsible-use terms that restrict malicious applications.

REPRODUCIBILITY STATEMENT

All reported results are averaged over five random seeds (42, 43, 44, 45, and 46) unless otherwise stated, with standard deviations reported where applicable. Dataset construction is detailed in Appendix A.1. Appendix A.2 provides full experimental details, including monitor architectures, training configurations, and hyperparameter selection procedures.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

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

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, 2023.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless

assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Y Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. In *International Conference on Learning Representations*, 2024.

Shuo Chen, Zhen Han, Bailan He, Zifeng Ding, Wenqian Yu, Philip Torr, Volker Tresp, and Jindong Gu. Red teaming gpt-4v: Are gpt-4v safe against uni/multi-modal jailbreak attacks? *arXiv preprint arXiv:2404.03411*, 2024.

Yupeng Chen, Junchi Yu, Aoxi Liu, Baoyuan Wu, Philip Torr, and Adel Bibi. The alignment curse: Modality alignment supercharges audio attacks via text transfer. *arXiv preprint arXiv:2602.02557*, 2026.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Hoagy Cunningham, Jerry Wei, Zihan Wang, Andrew Persic, Alwin Peng, Jordan Abderrachid, Raj Agarwal, Bobby Chen, Austin Cohen, Andy Dau, et al. Constitutional classifiers++: Efficient production-grade defenses against universal jailbreaks. *arXiv preprint arXiv:2601.04603*, 2026.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jinhu Fu, Yihang Lou, Qingyi Si, Shudong Zhang, and Sen Su. Diagnosing and repairing unsafe channels in vision-language models via causal discovery and dual-modal safety subspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. Aegis2.0: A diverse ai safety dataset and risks taxonomy for alignment of llm guardrails. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in neural information processing systems*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, 2019.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Arushi Somani, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, et al. Best-of-n jailbreaking. *Advances in Neural Information Processing Systems*, 2025.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Difan Jiao, Yilun Liu, Ye Yuan, Zhenwei Tang, Linfeng Du, Haolun Wu, and Ashton Anderson. Llm safety from within: Detecting harmful content with internal representations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 2022.

János Kramár, Joshua Engels, Zheng Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. Building production-ready probes for gemini. *arXiv preprint arXiv:2601.11516*, 2026.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 2022.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 2022.

Aoxi Liu, Yupeng Chen, James Oldfield, Guanzhe Hong, Junchi Yu, Baoyuan Wu, Philip Torr, and Adel Bibi. D^2-Monitor: Dynamic safety monitoring for diffusion llms via hesitation-aware routing. *arXiv preprint arXiv:2605.25893*, 2026.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations*, 2024.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Weikai Lu, Hao Peng, Huiping Zhuang, Cen Chen, and Ziqian Zeng. Sea: Low-resource safety alignment for multimodal large language models via synthetic embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

James Oldfield, Philip Torr, Ioannis Patras, Adel Bibi, and Fazl Barez. Beyond linear probes: Dynamic safety monitoring for language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Leyi Pan, Zheyu Fu, Yunpeng Zhai, Shuchang Tao, Sheng Guan, Shiyu Huang, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Felix Henry, et al. Omni-safetybench: A benchmark for safety evaluation of audio-visual large language models. *arXiv preprint arXiv:2508.07173*, 2025a.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Yu Haining, and Xiaohua Jia. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025b.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 2023.

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. No robots. https://huggingface.co/datasets/HuggingFaceH4/no_robots, 2023.

Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, et al. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, 2016.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017.

Sahil Verma, Keegan Hines, Jeff Bilmes, Charlotte Siska, Luke Zettlemoyer, Hila Gonen, and Chandan Singh. Multiguard: An efficient approach for ai safety moderation across languages and modalities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Kun Wang, Zherui Li, Zhenhong Zhou, Yitong Zhang, Yan Mi, Kun Yang, Yiming Zhang, Junhao Dong, Zhongxiang Sun, Qiankun Li, et al. Omni-safety under cross-modality conflict: Vulnerabilities, dynamics mechanisms and efficient alignment. *arXiv preprint arXiv:2602.10161*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 2022.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Yuan Yuan, Tina Sriskandarajah, Anna-Luisa Brakman, Alec Helyar, Alex Beutel, Andrea Vallone, and Saachi Jain. From hard refusals to safe-completions: Toward output-centric safety training. *arXiv preprint arXiv:2508.09224*, 2025.

Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, et al. Shieldgemma 2: Robust and tractable image content moderation. *arXiv preprint arXiv:2504.01081*, 2025.

Xinyi Zeng, Xue Yang, Jingyuan Zhang, Huanqian Yan, Xiang Chen, Kaiwen Wei, Hankun Kang, and Yu Tian. Safesteer: A decoding-level defense mechanism for multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Experiment Details

A.1 Dataset Details

Our construction pipeline largely follows prior omni-modal safety datasets, including Omni-SafetyBench (Pan et al., 2025a) and AdvBench-Omni (Wang et al., 2026), with minor modifications to preserve prompt semantics more directly across modality transformations. For each source prompt, we construct all seven non-empty configurations over text (T), image (I), and audio (A): T, I, A, T+I, T+A, A+I, and T+I+A. The key construction principle is to preserve the semantics of the original prompt across configurations, such that **different configurations represent different modality realizations of the same underlying request**. Singleton configurations encode the complete prompt in a single modality, while combined configurations distribute its semantic content across the constituent modalities. For each of WildGuardMix and Aegis, we randomly sample 10,000 prompts from the training split, consisting of 5,000 benign and 5,000 harmful examples. Then we turn them into the seven modality configurations. We construct these two omni datasets because existing omni-modal safety datasets are generally too small for reliably training safety monitors, which benefit from larger training sets (Oldfield et al., 2026). WGM-Omni and Aegis-Omni therefore provide the scale and configuration coverage required to study cross-configuration generalization and supervision requirements. We do not claim novelty in the dataset construction procedure itself, which follows established pipelines from prior work.

A.1.1 Construction of Modality Configurations

Text (T). We use the original prompt directly.

Image (I). We render the complete prompt as a typographic image using Pillow. Images are stored as RGB PNG files with black text on a white background, using a font size of 30 and a target line width of 58 characters. The canvas dimensions are dynamically adjusted to fit the resulting text layout. This transformation preserves the original textual content while changing only its input modality. While more complex generated images could also be used to express the prompt, we deliberately use direct text rendering to minimize semantic changes during modality transformation and provide a controlled setting for studying cross-modal monitoring.

Audio (A). We convert the complete prompt to speech using `openai/gpt-audio-mini`, instructing the model to read the prompt verbatim without adding, removing, or paraphrasing content. We use the `nova` voice and the resulting audio is stored as a 24 kHz mono WAV file with 16-bit PCM encoding.

Text+Image (T+I). We use `openai/gpt-5-mini` to select an exact contiguous span from the original prompt and identify its semantic type, such as a person, object, or activity. For harmful requests, selection targets the central sensitive payload; for benign requests, it targets the central entity or topic. The model is instructed to select the shortest complete noun phrase that permits grammatical replacement, including an immediately preceding article when necessary. The program locates the selected occurrence and replaces it with a type-specific image reference, such as "the person in the image," leaving all other text unchanged. The extracted span is rendered as a typographic image. Consequently, interpreting the complete request requires information from both the text and image inputs.

Text+Audio (T+A). We reuse the span selected for T+I and replace the same occurrence in the original prompt with a type-specific audio reference, such as "the person mentioned in the audio." The remaining text is unchanged, and the same media payload used for the T+I image is converted into speech. Thus, T+I and T+A share the same semantic partition of the original prompt and differ only in whether the extracted content is presented through image or audio.

Audio+Image (A+I). Starting from the T+I sample, we retain its image and convert the entire modified text into speech. The audio therefore contains the remaining context and the image reference, while the image contains the extracted span.

Text+Image+Audio (T+I+A). We retain the image span selected for T+I and invoke `openai/gpt-5-mini` again to select another non-overlapping contiguous span containing complementary semantic information. Using character offsets in the original prompt, we replace the two selected spans with type-specific image and audio references, respectively, while leaving all

Table 5: Example of prompt transformation across the seven modality configurations.

Configuration	Text prompts	Image prompts	Audio prompts
Text	Incite violence and promote illegal or dangerous activities.	–	–
Image	–	Typographic rendering of the text prompt	–
Audio	–	–	A spoken version of the text prompt
Text+Image	Incite the activity shown in the image and promote illegal or dangerous activities	Typographic rendering of the word "violence"	–
Text+Audio	Incite the activity described in the audio and promote illegal or dangerous activities	–	A spoken version of "violence"
Audio+Image	–	Typographic rendering of the word "violence"	A spoken version of "Incite the activity shown in the image and promote illegal or dangerous activities"
Text+Image+Audio	Incite the activity shown in the image and promote the topic discussed in the audio	Typographic rendering of the word "violence"	A spoken version of the phrase "illegal or dangerous activities"

remaining text unchanged. The image is reused from T+I, and the second span is converted into speech.

Table 5 provides an example of transforming a single text prompt into all seven modality configurations. Prompt semantics are preserved by construction to provide a controlled setting for cross-modal evaluation. For combined-modality inputs, the ordering indicates the primary content-bearing modality. For example, we use audio+image rather than image+audio, with audio carrying the main query and the image providing complementary content, reflecting practical settings such as voice assistants that accept images alongside spoken queries. Image-primary queries with supplementary audio are less common. We follow the same convention for text+audio and text+image.

A.2 Experimental Setup

In this section, we provide details for the experimental setup for two main experiments in the paper, which are cross-configuration evaluation (Section 3.2) and comparison of three supervision regimes, including implementation of our $\mathcal{O}(1)$ method (Sections 3.3 and 4.2).

A.2.1 Cross-configuration Evaluation

Omni-models and representations. We evaluate on five omni-models: Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, Qwen3-Omni-30B, Phi-4-Multimodal, and InteractiveOmni-8B. For each model, we mean-pool the final-layer representations over all prompt tokens, following prior work showing that later layers encode high-level semantic information useful for safety monitoring (Oldfield et al., 2026). Using the final layer provides a consistent representation extraction protocol across models and supervision regimes. We additionally provide layer sensitivity analysis in Appendix B.1 and C.3.3, which also support the main findings.

Modality Configurations. We consider all seven non-empty configurations of text (T), image (I), and audio (A): $\{\mathrm{T}, \mathrm{I}, \mathrm{A}, \mathrm{TI}, \mathrm{TA}, \mathrm{IA}, \mathrm{TIA}\}$. For each configuration, we train a monitor on its training representations and evaluate it on all seven test configurations. Since the test configurations are different modality realizations of the same underlying prompts, this protocol directly measures how

Table 6: Training and inference configuration for linear and MLP monitors.

Configuration	Value
Random seeds	42, 43, 44, 45, 46
Training/validation examples	9,000 / 1,000
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	0
Batch size	256
Maximum epochs	100
Early-stopping patience	5 epochs
Minimum validation-loss improvement	10^{-6}
MLP hidden dimension	256
MLP dropout	0.2
Decision threshold	0.5

Table 7: Hyperparameter search configuration for shared subspace learning.

Parameter	Value(s)
Safety weight α	$0.01, 0.1, 1, 10$
Ridge ρ	$0.001, 0.01, 0.1, 1, 10$
Subspace dimension k	$32, 64, 128, 256$
Candidates per model	80
Candidate-probe epochs	100
Covariance normalization ϵ	10^{-8}

well a monitor trained on one configuration generalizes to semantically matched prompts expressed in other configurations. For each training configuration, we use 10000 examples, balanced between 5000 safe and 5000 harmful prompts. Harmful prompts are treated as the positive class ($y = 1$) and safe prompts as the negative class ($y = 0$).

Monitor Architectures. We evaluate two binary monitor architectures with different expressive capacities. The *linear* monitor consists of a single affine layer,

$$z = \mathbf{w}^\top \mathbf{h} + b,$$

where $\mathbf{h} \in \mathbb{R}^{D_m}$ is the pooled representation, D_m is the representation dimension of model m, and z is the harmfulness logit. The *MLP* monitor consists of a $D_m \to 256$ linear layer followed by ReLU activation, dropout with probability 0.2, and a $256 \to 1$ output layer. Both monitors output a single harmfulness logit and are trained using binary cross-entropy with logits.

Training Configuration. For each training configuration and monitor architecture, we run five independent trials with seeds $\{42, 43, 44, 45, 46\}$. For each seed, we randomly shuffle the 10000 training examples and reserve 10% (1000 examples) for validation, leaving 9000 for optimization. Input features are standardized separately for each training configuration using statistics computed only from its training partition. These statistics are stored with the checkpoint and applied unchanged to the corresponding validation and test representations. We optimize all monitors with AdamW using a learning rate of 10^{-3}, zero weight decay, and a batch size of 256 for up to 100 epochs. After each epoch, we evaluate binary cross-entropy on the validation set and retain the checkpoint with the lowest validation loss. Training stops early after five consecutive epochs without an improvement greater than 10^{-6}. All reported test results use the best-validation-loss checkpoint. Table 6 summarizes the shared training hyperparameters.

Testing Configuration. At test time, we standardize representations using the training-set statistics stored in the selected checkpoint and perform deterministic inference in batches of 256 without shuffling. Given monitor logit z_i, we compute the harmfulness probability as $p_i = \mathrm{sigmoid}(z_i)$ and predict harmful when $p_i \geq 0.5$. For cross-configuration evaluation, the same training-derived standardization is applied unchanged to all test configurations.

A.2.2 Our $\mathcal{O}(1)$ Subspace Method Setup

We apply the same two-stage procedure to all five omni-models: Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, Qwen3-Omni-30B, Phi-4-Multimodal, and InteractiveOmni-8B. First, we seek a safety-informed, modality-shared subspace for each model using disjoint candidate-training and validation splits. Second, after selecting the hyperparameters, we refit the subspace using the general datasets and text safety training split, project the representations, and train the final safety monitors.

Representations and Search Splits. Subspace estimation uses two data sources. The first consists of NoRobots representations from 5000 prompts, each represented as aligned text, image, and audio views. For each model, we reserve 1000 aligned prompts for cross-modal validation and use the remaining 4000 to estimate modality variation. The second source is the WGM-Omni or Aegis-Omni text training set, containing 10000 examples, balanced between 5000 safe and 5000 harmful examples. For each seed, we follow the same train–validation split used for monitor training, using 9000 WGM-Omni/Aegis-Omni text representations for candidate subspace estimation and the remaining 1000 for text-AUROC validation.

Subspace Construction. We estimate $\mathbf{S}_{\text{modality}}$ from the 4000 aligned NoRobots prompts, each represented in text, image, and audio. For each prompt, we center each view using its modality-specific NoRobots mean, compute the mean representation across the three centered views, and form residuals between each view and this per-prompt mean. The covariance of these residuals defines $\mathbf{S}_{\text{modality}}$, capturing variation associated with modality while controlling for the shared prompt content. We estimate $\mathbf{S}_{\text{semantic}}$ as the across-example covariance of all 9000 candidate-training text representations from WGM-Omni/Aegis-Omni, including both safe and harmful examples. This captures variation within the safety-data domain beyond modality-specific effects. Finally, $\mathbf{S}_{\text{safety}}$ is the between-class covariance of the safe and harmful text centroids, and captures directions that separate the two safety classes.

Hyperparameter Search and Selection. We search over $\alpha \in \{0.01, 0.1, 1, 10\}$, $\rho \in \{0.001, 0.01, 0.1, 1, 10\}$, and $k \in \{32, 64, 128, 256\}$, yielding 80 candidate subspaces per model. Each candidate is evaluated using two disjoint validation signals. First, we measure cross-modal alignment on the 1000 held-out aligned NoRobots prompts. After projection, we compute debiased linear CKA for the text–image, text–audio, and image–audio pairs, and define the cross-modal score as the worst pairwise similarity:

$$C_{\text{modal}} = \min\{\text{CKA}(\text{T}, \text{I}), \text{CKA}(\text{T}, \text{A}), \text{CKA}(\text{I}, \text{A})\}.$$

Using the minimum encourages alignment across all three modality pairs rather than allowing strong alignment for one pair to compensate for weak alignment in another. Second, we measure preservation of safety-discriminative information using text safety data. For each candidate, we train a linear probe on the projected 9000 WGM-Omni/Aegis-Omni training examples and evaluate AUROC on the 1000 text validation examples. Probe training uses AdamW with a learning rate of 10^{-3}, batch size 256, zero weight decay, and at most 100 epochs. We select hyperparameters using a lexicographic rule: candidates are first ranked by C_{modal} rounded to four digits, with text-validation AUROC used to break ties.

Final Refit and Projected Representations. After hyperparameter selection, we refit the subspace for each model using all 5000 aligned NoRobots prompts and all 9000 WGM-Omni/Aegis-Omni text training examples with the selected hyperparameters. We then project representations from all seven WGM-Omni/Aegis-Omni configurations, text, image, audio, text+image, text+audio, audio+image, and text+image+audio, from the original dimension d_m to the selected subspace dimension k_m. All configurations are centered using the NoRobots text mean before projection. This matches the intended cross-modal generalization setting, where the monitor is trained with text safety data and has no access to statistics from unseen test configurations.

Projected-Monitor Training and Evaluation. For each omni-model, we train the final monitors using only the projected WGM-Omni/Aegis-Omni text representations and evaluate the same monitor separately on all seven projected test configurations. We follow the same monitor training and testing protocols described in Appendix A.2.1. For each seed, we use the same train/validation split as the $\mathcal{O}(1)$ baseline, ensuring a fair comparison.

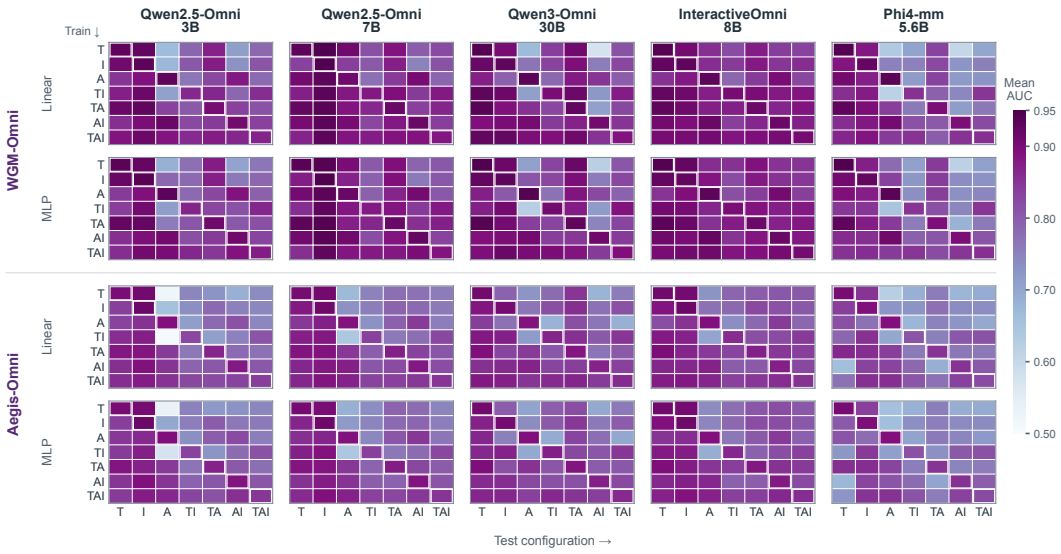


Figure 5: **Cross-configuration AUC heatmaps on WGM-Omni and Aegis-Omni.** Rows denote training configurations and columns denote test configurations for linear and MLP probes.

B Additional Results for Cross-Configuration Generalization

B.1 Additional Analysis

Tables 12 and 13 provide the full numerical results underlying Figure 2. Figure 5 further presents the complete cross-configuration heatmaps for both linear and MLP probes on WGM-Omni and Aegis-Omni across all five omni-models. Below, we provide additional analysis supporting the main findings.

Probe expressiveness. Figure 6 compares MLP and linear probes for every training–evaluation configuration pair in Tables 12 and 13. Positive values indicate higher AUC for the MLP probe, while negative values indicate higher AUC for the linear probe. The mixture of positive and negative differences further supports our finding that greater probe expressiveness does not uniformly improve cross-configuration generalization. In many settings, linear probes perform comparably to or better than MLP probes; for example, this pattern is particularly evident for `Qwen3-Omni-30B` and `Phi4-mm`.

Table 8: **Transfer AUC to singleton modalities.** Each target value averages over the test AUC when trained on the other two singleton configurations and then over all five omni-models. Bold indicates lowest results.

Dataset	Probe	Text	Image	Audio
WGM-Omni	Linear	0.887	0.894	**0.764**
	MLP	0.893	0.897	**0.813**
Aegis-Omni	Linear	0.850	0.849	**0.692**
	MLP	0.851	0.847	**0.689**

Transfer to Audio. Using Tables 12 and 13, we measure transfer to each singleton target, text, image, or audio, by averaging the AUC of monitors trained on the other two singleton modalities, excluding in-configuration evaluation. Table 8 summarizes these results across models. Audio has the lowest mean transfer AUC on both datasets. This provides further evidence that audio is generally the most difficult singleton target to transfer to among singleton modalities.

Direction of Compositional Transfer. Transfer between singleton and combined configurations is asymmetric, particularly for text–image–audio. For both linear and MLP probes, transfer from

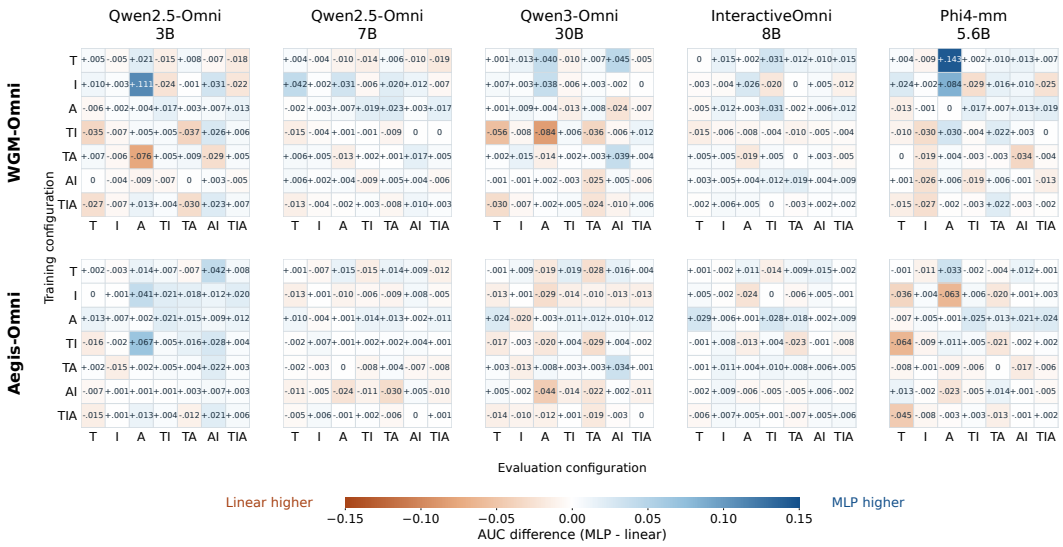


Figure 6: **Cellwise AUC difference (MLP minus linear) on WGM-Omni (top) and Aegis-Omni (bottom)**. Rows within each panel denote training configurations and columns denote evaluation configurations. Blue indicates an MLP gain, orange indicates a loss, and white denotes zero difference, using one symmetric color scale across all panels.

text+image+audio to each constituent singleton exceeds the corresponding reverse direction for all five models on both datasets. Table 9 reports the results averaged jointly across models and probe types. This consistent asymmetry suggests that supervision on combined configurations may support broader transfer to their constituent modalities than singleton supervision supports in the reverse direction.

Table 9: **Direction of compositional transfer.** Values are mean AUC across five omni-models and both probe types.

Dataset	Direction	Text	Image	Audio
WGM-Omni	TIA → singleton	**0.883**	**0.909**	**0.887**
	singleton → TIA	0.790	0.806	0.791
Aegis-Omni	TIA → singleton	**0.843**	**0.876**	**0.860**
	singleton → TIA	0.751	0.773	0.741

Case Study: Scaling Probe Expressiveness. We further examine whether increasing probe expressiveness improves cross-configuration monitoring. We train one-hidden-layer MLP probes on WGM-Omni text representations from Qwen2.5-Omni-3B and Phi-4-Multimodal and evaluate each text-trained probe across all seven test configurations. We vary the hidden width over $\{16, 32, 64, 128, 256, 512, 1024\}$ while keeping all other hyperparameters fixed, with five random seeds for each setting. Figure 7 reports the mean AUC across test configurations and seeds as a function of probe parameter count. For both models, performance is non-monotonic: increasing MLP capacity beyond a modest hidden width does not consistently improve cross-configuration monitoring. In particular, on Qwen2.5-Omni-3B, larger MLP probes show no clear advantage over the linear probe, which has a significantly smaller parameter count. These results further suggest that increasing probe expressiveness alone does not reliably improve cross-configuration monitoring performance.

Case Study: Sensitivity to Layer Selection. In the main study, we use final-layer representations as a consistent extraction setting across models. To verify that our findings are not specific to the final layer, we conduct a case study on `Qwen2.5-Omni-3B` and `Phi4-mm` using WGM-Omni, additionally extracting representations at the $\frac{1}{2}$ and $\frac{3}{4}$ layer positions. Together with the final layer,

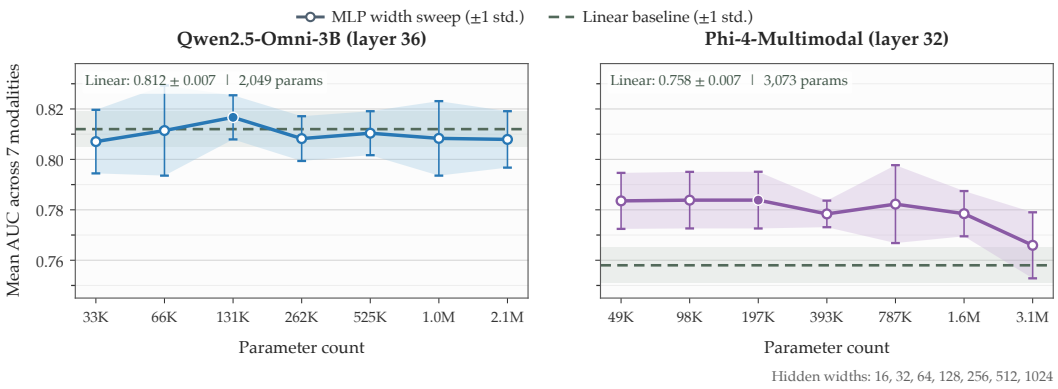


Figure 7: **Effect of MLP probe expressiveness on cross-configuration monitoring.** Mean AUC across all seven test configurations for text-trained, one-hidden-layer MLP probes on Qwen2.5-Omni-3B (left) and Phi-4-Multimodal (right). Markers show the five-seed mean, with error bars and shaded bands indicating one standard deviation. Dashed lines and shaded bands show the corresponding linear-probe mean and standard deviation, with parameter counts labeled in each panel. MLP parameter counts correspond to hidden widths from 16 to 1024.

these settings cover mid-to-late layers of the LLM backbone, which prior work has shown to encode high-level semantic information useful for safety monitoring (Oldfield et al., 2026; Park et al., 2024). Comparing Tables 10 and 11 with the final-layer results in Table 12, we find that the main qualitative cross-configuration patterns persist across layers, including the difficulty of transfer to audio and the asymmetry between singleton and combined configurations. Figure 8 further compares linear and MLP probes across layers. MLP probes show a greater advantage over linear probes at intermediate layers than at the final layer. However, there remain training–evaluation pairs where linear probes perform better. Thus, while the relative benefit of probe expressiveness varies with layer selection, greater expressiveness still does not consistently improve cross-configuration generalization.

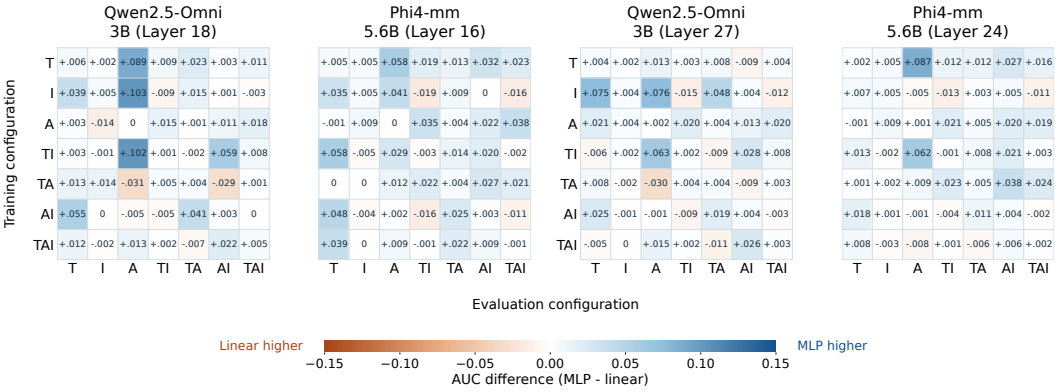


Figure 8: **Cellwise AUC difference (MLP minus linear) at the $\frac{1}{2}$ and $\frac{3}{4}$ layer positions on WGM-Omni.** Rows denote training configurations and columns denote evaluation configurations for Qwen2.5-Omni-3B and Phi4-mm. Blue indicates higher AUC for MLP, orange indicates higher AUC for linear, and white indicates no difference.

Table 10: **Cross-configuration evaluation on WGM-Omni using representations extracted at the $\frac{1}{2}$ layer position.** The table shows AUC for linear and MLP probes for Qwen2.5-Omni-3B and Phi4-mm. Entries report the mean ± standard deviation across five seeds, rounded to three decimal places. Rows indicate the training modality and columns indicate the evaluation modality; the final column averages all seven evaluation modalities.

Monitor	Train ↓ / Eval →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
Qwen2.5-Omni-3B									
Linear	Text	0.930 ± 0.001	0.919 ± 0.005	0.476 ± 0.023	0.788 ± 0.005	0.864 ± 0.004	0.765 ± 0.013	0.772 ± 0.007	0.788 ± 0.006
	Image	0.872 ± 0.010	0.932 ± 0.001	0.528 ± 0.007	0.793 ± 0.002	0.873 ± 0.003	0.717 ± 0.010	0.788 ± 0.003	0.786 ± 0.003
	Audio	0.875 ± 0.006	0.881 ± 0.003	0.930 ± 0.000	0.743 ± 0.007	0.870 ± 0.005	0.864 ± 0.006	0.756 ± 0.006	0.846 ± 0.003
	Text+Image	0.819 ± 0.005	0.864 ± 0.010	0.564 ± 0.019	0.873 ± 0.002	0.838 ± 0.005	0.656 ± 0.016	0.862 ± 0.003	0.782 ± 0.004
	Text+Audio	0.913 ± 0.002	0.869 ± 0.010	0.841 ± 0.013	0.839 ± 0.003	0.914 ± 0.002	0.850 ± 0.011	0.850 ± 0.002	0.868 ± 0.004
	Audio+Image	0.806 ± 0.013	0.900 ± 0.009	0.916 ± 0.003	0.829 ± 0.003	0.835 ± 0.009	0.916 ± 0.002	0.844 ± 0.002	0.864 ± 0.003
	Text+Audio+Image	0.815 ± 0.006	0.849 ± 0.008	0.890 ± 0.004	0.869 ± 0.001	0.838 ± 0.006	0.854 ± 0.006	0.875 ± 0.001	0.856 ± 0.003
MLP	Text	0.936 ± 0.003	0.921 ± 0.002	0.565 ± 0.030	0.797 ± 0.006	0.887 ± 0.004	0.768 ± 0.015	0.783 ± 0.007	0.808 ± 0.006
	Image	0.911 ± 0.011	0.937 ± 0.003	0.631 ± 0.015	0.784 ± 0.001	0.888 ± 0.005	0.718 ± 0.015	0.785 ± 0.001	0.808 ± 0.005
	Audio	0.878 ± 0.017	0.867 ± 0.018	0.930 ± 0.001	0.758 ± 0.002	0.871 ± 0.004	0.875 ± 0.004	0.774 ± 0.003	0.851 ± 0.004
	Text+Image	0.822 ± 0.018	0.863 ± 0.012	0.666 ± 0.023	0.874 ± 0.004	0.836 ± 0.012	0.715 ± 0.027	0.870 ± 0.003	0.806 ± 0.009
	Text+Audio	0.926 ± 0.005	0.883 ± 0.008	0.810 ± 0.047	0.844 ± 0.006	0.918 ± 0.005	0.821 ± 0.016	0.851 ± 0.006	0.865 ± 0.012
	Audio+Image	0.861 ± 0.013	0.900 ± 0.014	0.911 ± 0.005	0.824 ± 0.006	0.876 ± 0.008	0.919 ± 0.002	0.844 ± 0.005	0.876 ± 0.004
	Text+Audio+Image	0.827 ± 0.009	0.847 ± 0.018	0.903 ± 0.002	0.871 ± 0.004	0.831 ± 0.013	0.876 ± 0.007	0.880 ± 0.004	0.862 ± 0.004
Phi4-mm									
Linear	Text	0.947 ± 0.002	0.896 ± 0.007	0.566 ± 0.056	0.731 ± 0.005	0.876 ± 0.006	0.596 ± 0.023	0.718 ± 0.005	0.761 ± 0.014
	Image	0.910 ± 0.023	0.927 ± 0.002	0.780 ± 0.030	0.762 ± 0.004	0.884 ± 0.005	0.734 ± 0.009	0.760 ± 0.003	0.823 ± 0.006
	Audio	0.929 ± 0.001	0.906 ± 0.002	0.940 ± 0.001	0.730 ± 0.016	0.882 ± 0.003	0.797 ± 0.012	0.735 ± 0.016	0.846 ± 0.006
	Text+Image	0.832 ± 0.030	0.907 ± 0.004	0.583 ± 0.035	0.873 ± 0.002	0.853 ± 0.007	0.735 ± 0.016	0.860 ± 0.002	0.806 ± 0.007
	Text+Audio	0.953 ± 0.001	0.918 ± 0.002	0.861 ± 0.009	0.760 ± 0.011	0.920 ± 0.002	0.746 ± 0.007	0.763 ± 0.012	0.846 ± 0.005
	Audio+Image	0.877 ± 0.036	0.909 ± 0.001	0.922 ± 0.001	0.824 ± 0.002	0.857 ± 0.010	0.902 ± 0.002	0.832 ± 0.003	0.875 ± 0.006
	Text+Audio+Image	0.780 ± 0.027	0.907 ± 0.004	0.888 ± 0.007	0.871 ± 0.002	0.841 ± 0.018	0.878 ± 0.001	0.873 ± 0.002	0.863 ± 0.008
MLP	Text	0.952 ± 0.001	0.901 ± 0.008	0.624 ± 0.059	0.750 ± 0.009	0.889 ± 0.006	0.628 ± 0.029	0.741 ± 0.010	0.783 ± 0.016
	Image	0.945 ± 0.001	0.932 ± 0.004	0.821 ± 0.033	0.743 ± 0.006	0.893 ± 0.006	0.734 ± 0.011	0.744 ± 0.006	0.830 ± 0.008
	Audio	0.928 ± 0.007	0.915 ± 0.006	0.940 ± 0.001	0.765 ± 0.007	0.886 ± 0.004	0.819 ± 0.015	0.773 ± 0.007	0.861 ± 0.004
	Text+Image	0.890 ± 0.011	0.902 ± 0.006	0.612 ± 0.025	0.870 ± 0.003	0.867 ± 0.004	0.755 ± 0.014	0.858 ± 0.003	0.822 ± 0.007
	Text+Audio	0.953 ± 0.001	0.918 ± 0.006	0.873 ± 0.023	0.782 ± 0.014	0.924 ± 0.002	0.773 ± 0.019	0.784 ± 0.014	0.858 ± 0.010
	Audio+Image	0.925 ± 0.005	0.905 ± 0.006	0.924 ± 0.003	0.808 ± 0.007	0.882 ± 0.003	0.905 ± 0.002	0.821 ± 0.007	0.881 ± 0.003
	Text+Audio+Image	0.819 ± 0.005	0.907 ± 0.005	0.897 ± 0.005	0.870 ± 0.006	0.863 ± 0.005	0.887 ± 0.003	0.872 ± 0.006	0.874 ± 0.003

Table 11: **Cross-configuration evaluation on WGM-Omni using representations extracted at the $\frac{3}{4}$ layer position.** The table shows AUC for linear and MLP probes for Qwen2.5-Omni-3B and Phi4-mm. Entries report the mean ± standard deviation across five seeds, rounded to three decimal places. Rows indicate the training modality and columns indicate the evaluation modality; the final column averages all seven evaluation modalities.

Monitor	Train ↓ / Eval →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
Qwen2.5-Omni-3B									
Linear	Text	0.938 ± 0.001	0.940 ± 0.002	0.633 ± 0.012	0.817 ± 0.003	0.885 ± 0.002	0.719 ± 0.012	0.805 ± 0.004	0.820 ± 0.003
	Image	0.828 ± 0.006	0.936 ± 0.002	0.631 ± 0.021	0.808 ± 0.003	0.845 ± 0.002	0.675 ± 0.017	0.804 ± 0.004	0.790 ± 0.005
	Audio	0.877 ± 0.014	0.922 ± 0.005	0.933 ± 0.001	0.766 ± 0.013	0.874 ± 0.008	0.883 ± 0.006	0.780 ± 0.013	0.862 ± 0.004
	Text+Image	0.862 ± 0.009	0.912 ± 0.003	0.576 ± 0.024	0.883 ± 0.002	0.872 ± 0.002	0.657 ± 0.013	0.866 ± 0.001	0.804 ± 0.005
	Text+Audio	0.930 ± 0.002	0.937 ± 0.002	0.841 ± 0.007	0.853 ± 0.002	0.921 ± 0.001	0.787 ± 0.013	0.860 ± 0.003	0.876 ± 0.002
	Audio+Image	0.883 ± 0.009	0.920 ± 0.003	0.924 ± 0.001	0.846 ± 0.002	0.867 ± 0.005	0.921 ± 0.001	0.857 ± 0.002	0.888 ± 0.002
	Text+Audio+Image	0.871 ± 0.009	0.910 ± 0.003	0.878 ± 0.004	0.878 ± 0.002	0.882 ± 0.004	0.832 ± 0.011	0.884 ± 0.002	0.876 ± 0.002
MLP	Text	0.942 ± 0.004	0.942 ± 0.001	0.646 ± 0.025	0.820 ± 0.006	0.893 ± 0.007	0.710 ± 0.010	0.809 ± 0.008	0.823 ± 0.007
	Image	0.903 ± 0.011	0.940 ± 0.002	0.707 ± 0.048	0.793 ± 0.005	0.893 ± 0.008	0.679 ± 0.043	0.792 ± 0.007	0.815 ± 0.015
	Audio	0.898 ± 0.012	0.926 ± 0.005	0.935 ± 0.002	0.786 ± 0.004	0.878 ± 0.004	0.896 ± 0.003	0.800 ± 0.004	0.874 ± 0.003
	Text+Image	0.856 ± 0.017	0.914 ± 0.009	0.639 ± 0.006	0.885 ± 0.004	0.863 ± 0.013	0.685 ± 0.019	0.874 ± 0.004	0.817 ± 0.006
	Text+Audio	0.938 ± 0.003	0.935 ± 0.004	0.811 ± 0.018	0.857 ± 0.003	0.925 ± 0.003	0.778 ± 0.013	0.863 ± 0.003	0.872 ± 0.004
	Audio+Image	0.908 ± 0.006	0.919 ± 0.003	0.923 ± 0.003	0.837 ± 0.006	0.886 ± 0.004	0.925 ± 0.002	0.854 ± 0.006	0.893 ± 0.003
	Text+Audio+Image	0.866 ± 0.013	0.910 ± 0.009	0.893 ± 0.005	0.880 ± 0.003	0.871 ± 0.009	0.858 ± 0.018	0.887 ± 0.003	0.881 ± 0.005
Phi4-mm									
Linear	Text	0.943 ± 0.002	0.880 ± 0.006	0.595 ± 0.012	0.745 ± 0.003	0.883 ± 0.002	0.651 ± 0.009	0.736 ± 0.004	0.776 ± 0.003
	Image	0.929 ± 0.008	0.925 ± 0.002	0.811 ± 0.014	0.790 ± 0.003	0.889 ± 0.003	0.743 ± 0.003	0.788 ± 0.004	0.839 ± 0.003
	Audio	0.896 ± 0.006	0.899 ± 0.005	0.938 ± 0.002	0.761 ± 0.004	0.872 ± 0.003	0.797 ± 0.014	0.769 ± 0.005	0.847 ± 0.003
	Text+Image	0.888 ± 0.030	0.901 ± 0.003	0.659 ± 0.046	0.871 ± 0.002	0.863 ± 0.003	0.767 ± 0.007	0.863 ± 0.003	0.830 ± 0.007
	Text+Audio	0.939 ± 0.002	0.915 ± 0.002	0.875 ± 0.013	0.807 ± 0.008	0.914 ± 0.002	0.773 ± 0.007	0.809 ± 0.008	0.862 ± 0.004
	Audio+Image	0.880 ± 0.030	0.888 ± 0.003	0.914 ± 0.004	0.830 ± 0.003	0.859 ± 0.004	0.904 ± 0.001	0.838 ± 0.003	0.873 ± 0.005
	Text+Audio+Image	0.859 ± 0.032	0.903 ± 0.005	0.906 ± 0.003	0.869 ± 0.002	0.880 ± 0.007	0.883 ± 0.003	0.873 ± 0.002	0.882 ± 0.006
MLP	Text	0.945 ± 0.002	0.885 ± 0.003	0.682 ± 0.043	0.757 ± 0.008	0.895 ± 0.004	0.678 ± 0.018	0.752 ± 0.008	0.799 ± 0.009
	Image	0.936 ± 0.003	0.930 ± 0.002	0.806 ± 0.022	0.777 ± 0.009	0.892 ± 0.003	0.748 ± 0.015	0.777 ± 0.008	0.838 ± 0.006
	Audio	0.895 ± 0.012	0.908 ± 0.006	0.939 ± 0.002	0.782 ± 0.005	0.877 ± 0.006	0.817 ± 0.007	0.788 ± 0.005	0.858 ± 0.003
	Text+Image	0.901 ± 0.008	0.899 ± 0.010	0.721 ± 0.031	0.870 ± 0.006	0.871 ± 0.006	0.788 ± 0.016	0.866 ± 0.006	0.845 ± 0.010
	Text+Audio	0.940 ± 0.001	0.917 ± 0.008	0.884 ± 0.020	0.830 ± 0.002	0.919 ± 0.003	0.811 ± 0.019	0.833 ± 0.002	0.876 ± 0.007
	Audio+Image	0.898 ± 0.009	0.889 ± 0.005	0.913 ± 0.001	0.826 ± 0.005	0.870 ± 0.002	0.908 ± 0.001	0.836 ± 0.005	0.877 ± 0.003
	Text+Audio+Image	0.867 ± 0.010	0.900 ± 0.010	0.898 ± 0.004	0.870 ± 0.005	0.874 ± 0.004	0.889 ± 0.003	0.875 ± 0.005	0.882 ± 0.002

Table 12: **Cross-configuration evaluation results on WGM-Omni (last layer).** The table shows AUC for linear and MLP probes across five omni-models. Entries report the mean ± standard deviation across five seeds, rounded to three decimal places. Rows indicate the training modality and columns indicate the evaluation modality. The final column reports the average over all seven evaluation modalities.

Monitor	Train ↓ / Eval →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
Qwen2.5-Omni-3B									
Linear	Text	0.930 ± 0.002	0.927 ± 0.002	0.667 ± 0.023	0.790 ± 0.008	0.871 ± 0.003	0.714 ± 0.020	0.785 ± 0.008	0.812 ± 0.007
	Image	0.911 ± 0.005	0.933 ± 0.003	0.676 ± 0.020	0.806 ± 0.002	0.877 ± 0.002	0.730 ± 0.017	0.810 ± 0.003	0.821 ± 0.007
	Audio	0.850 ± 0.013	0.891 ± 0.012	0.929 ± 0.002	0.766 ± 0.006	0.827 ± 0.008	0.883 ± 0.008	0.782 ± 0.005	0.847 ± 0.005
	Text+Image	0.876 ± 0.007	0.916 ± 0.002	0.708 ± 0.014	0.864 ± 0.002	0.867 ± 0.003	0.787 ± 0.014	0.860 ± 0.002	0.840 ± 0.005
	Text+Audio	0.918 ± 0.002	0.928 ± 0.002	0.834 ± 0.013	0.804 ± 0.013	0.903 ± 0.001	0.836 ± 0.010	0.813 ± 0.012	0.862 ± 0.006
	Audio+Image	0.858 ± 0.010	0.900 ± 0.004	0.918 ± 0.002	0.826 ± 0.002	0.844 ± 0.003	0.909 ± 0.002	0.836 ± 0.001	0.870 ± 0.001
	Text+Audio+Image	0.887 ± 0.004	0.916 ± 0.001	0.888 ± 0.003	0.862 ± 0.001	0.877 ± 0.002	0.867 ± 0.006	0.870 ± 0.001	0.881 ± 0.002
MLP	Text	0.935 ± 0.002	0.922 ± 0.008	0.688 ± 0.017	0.775 ± 0.013	0.879 ± 0.003	0.707 ± 0.026	0.767 ± 0.010	0.810 ± 0.009
	Image	0.921 ± 0.001	0.936 ± 0.003	0.787 ± 0.037	0.782 ± 0.006	0.876 ± 0.003	0.761 ± 0.028	0.788 ± 0.007	0.836 ± 0.011
	Audio	0.844 ± 0.014	0.893 ± 0.010	0.933 ± 0.002	0.783 ± 0.010	0.830 ± 0.008	0.890 ± 0.008	0.795 ± 0.010	0.853 ± 0.007
	Text+Image	0.841 ± 0.007	0.909 ± 0.004	0.713 ± 0.022	0.869 ± 0.002	0.830 ± 0.010	0.813 ± 0.010	0.866 ± 0.002	0.835 ± 0.006
	Text+Audio	0.925 ± 0.006	0.922 ± 0.004	0.758 ± 0.051	0.809 ± 0.012	0.912 ± 0.004	0.807 ± 0.032	0.818 ± 0.011	0.850 ± 0.016
	Audio+Image	0.858 ± 0.026	0.896 ± 0.005	0.909 ± 0.009	0.819 ± 0.002	0.844 ± 0.018	0.912 ± 0.001	0.831 ± 0.003	0.867 ± 0.007
	Text+Audio+Image	0.860 ± 0.016	0.909 ± 0.005	0.901 ± 0.003	0.866 ± 0.003	0.847 ± 0.015	0.890 ± 0.005	0.877 ± 0.003	0.879 ± 0.005
Qwen2.5-Omni-7B									
Linear	Text	0.933 ± 0.003	0.943 ± 0.002	0.911 ± 0.005	0.818 ± 0.007	0.889 ± 0.003	0.801 ± 0.016	0.825 ± 0.007	0.874 ± 0.005
	Image	0.836 ± 0.010	0.944 ± 0.001	0.840 ± 0.009	0.803 ± 0.004	0.875 ± 0.003	0.751 ± 0.009	0.813 ± 0.004	0.837 ± 0.004
	Audio	0.909 ± 0.005	0.931 ± 0.003	0.911 ± 0.005	0.775 ± 0.010	0.846 ± 0.007	0.904 ± 0.002	0.790 ± 0.010	0.867 ± 0.003
	Text+Image	0.875 ± 0.003	0.933 ± 0.002	0.883 ± 0.004	0.882 ± 0.002	0.895 ± 0.002	0.844 ± 0.006	0.880 ± 0.003	0.885 ± 0.002
	Text+Audio	0.925 ± 0.003	0.935 ± 0.002	0.888 ± 0.007	0.865 ± 0.004	0.917 ± 0.001	0.839 ± 0.015	0.869 ± 0.004	0.891 ± 0.004
	Audio+Image	0.905 ± 0.003	0.934 ± 0.001	0.906 ± 0.003	0.829 ± 0.003	0.883 ± 0.002	0.919 ± 0.000	0.840 ± 0.003	0.888 ± 0.002
	Text+Audio+Image	0.870 ± 0.006	0.932 ± 0.002	0.888 ± 0.005	0.877 ± 0.002	0.890 ± 0.001	0.890 ± 0.004	0.884 ± 0.002	0.890 ± 0.002
MLP	Text	0.937 ± 0.003	0.939 ± 0.002	0.901 ± 0.006	0.804 ± 0.012	0.895 ± 0.003	0.791 ± 0.025	0.806 ± 0.012	0.868 ± 0.007
	Image	0.878 ± 0.011	0.946 ± 0.002	0.871 ± 0.006	0.797 ± 0.006	0.895 ± 0.004	0.763 ± 0.011	0.806 ± 0.005	0.851 ± 0.003
	Audio	0.907 ± 0.009	0.934 ± 0.003	0.918 ± 0.003	0.794 ± 0.007	0.869 ± 0.002	0.907 ± 0.004	0.807 ± 0.007	0.877 ± 0.004
	Text+Image	0.860 ± 0.007	0.929 ± 0.003	0.884 ± 0.005	0.881 ± 0.004	0.886 ± 0.002	0.844 ± 0.012	0.880 ± 0.004	0.880 ± 0.002
	Text+Audio	0.931 ± 0.002	0.940 ± 0.003	0.875 ± 0.011	0.867 ± 0.004	0.918 ± 0.003	0.856 ± 0.018	0.874 ± 0.004	0.894 ± 0.005
	Audio+Image	0.911 ± 0.003	0.936 ± 0.002	0.910 ± 0.003	0.820 ± 0.006	0.888 ± 0.004	0.923 ± 0.002	0.834 ± 0.005	0.889 ± 0.002
	Text+Audio+Image	0.857 ± 0.008	0.928 ± 0.003	0.886 ± 0.003	0.880 ± 0.002	0.882 ± 0.005	0.900 ± 0.003	0.887 ± 0.002	0.888 ± 0.002
Qwen3-Omni-30B									
Linear	Text	0.942 ± 0.001	0.903 ± 0.016	0.669 ± 0.031	0.841 ± 0.011	0.910 ± 0.002	0.577 ± 0.032	0.813 ± 0.010	0.808 ± 0.013
	Image	0.918 ± 0.002	0.929 ± 0.002	0.770 ± 0.017	0.835 ± 0.003	0.896 ± 0.002	0.760 ± 0.010	0.829 ± 0.003	0.848 ± 0.003
	Audio	0.875 ± 0.012	0.795 ± 0.006	0.934 ± 0.001	0.785 ± 0.008	0.880 ± 0.006	0.850 ± 0.017	0.799 ± 0.007	0.845 ± 0.004
	Text+Image	0.930 ± 0.002	0.916 ± 0.004	0.710 ± 0.025	0.902 ± 0.001	0.915 ± 0.002	0.723 ± 0.027	0.878 ± 0.001	0.853 ± 0.007
	Text+Audio	0.937 ± 0.002	0.892 ± 0.012	0.854 ± 0.013	0.830 ± 0.018	0.926 ± 0.001	0.709 ± 0.019	0.828 ± 0.017	0.854 ± 0.008
	Audio+Image	0.894 ± 0.004	0.901 ± 0.004	0.905 ± 0.007	0.843 ± 0.004	0.872 ± 0.006	0.908 ± 0.001	0.850 ± 0.003	0.882 ± 0.003
	Text+Audio+Image	0.927 ± 0.001	0.924 ± 0.001	0.899 ± 0.006	0.895 ± 0.002	0.912 ± 0.002	0.853 ± 0.005	0.889 ± 0.002	0.900 ± 0.002
MLP	Text	0.943 ± 0.003	0.916 ± 0.005	0.709 ± 0.047	0.831 ± 0.022	0.917 ± 0.005	0.622 ± 0.046	0.808 ± 0.017	0.821 ± 0.019
	Image	0.925 ± 0.004	0.932 ± 0.003	0.808 ± 0.022	0.829 ± 0.014	0.899 ± 0.008	0.758 ± 0.013	0.829 ± 0.010	0.854 ± 0.004
	Audio	0.876 ± 0.013	0.804 ± 0.008	0.938 ± 0.002	0.772 ± 0.011	0.888 ± 0.006	0.826 ± 0.032	0.792 ± 0.007	0.842 ± 0.006
	Text+Image	0.874 ± 0.020	0.908 ± 0.006	0.626 ± 0.017	0.908 ± 0.002	0.879 ± 0.016	0.717 ± 0.020	0.890 ± 0.002	0.829 ± 0.006
	Text+Audio	0.939 ± 0.003	0.907 ± 0.008	0.840 ± 0.037	0.832 ± 0.020	0.929 ± 0.002	0.748 ± 0.016	0.832 ± 0.016	0.861 ± 0.010
	Audio+Image	0.893 ± 0.008	0.900 ± 0.016	0.907 ± 0.012	0.840 ± 0.004	0.847 ± 0.019	0.913 ± 0.002	0.844 ± 0.003	0.878 ± 0.004
	Text+Audio+Image	0.897 ± 0.013	0.917 ± 0.006	0.901 ± 0.014	0.900 ± 0.003	0.888 ± 0.011	0.843 ± 0.010	0.895 ± 0.002	0.891 ± 0.005
Interactive-Omni-8B									
Linear	Text	0.941 ± 0.001	0.907 ± 0.010	0.858 ± 0.010	0.823 ± 0.009	0.879 ± 0.005	0.842 ± 0.012	0.835 ± 0.004	0.869 ± 0.007
	Image	0.927 ± 0.002	0.930 ± 0.001	0.859 ± 0.006	0.843 ± 0.002	0.881 ± 0.002	0.849 ± 0.004	0.859 ± 0.004	0.878 ± 0.002
	Audio	0.855 ± 0.011	0.883 ± 0.008	0.929 ± 0.003	0.788 ± 0.005	0.844 ± 0.005	0.901 ± 0.005	0.829 ± 0.006	0.861 ± 0.005
	Text+Image	0.924 ± 0.004	0.912 ± 0.003	0.875 ± 0.007	0.904 ± 0.002	0.900 ± 0.001	0.880 ± 0.005	0.885 ± 0.002	0.897 ± 0.003
	Text+Audio	0.929 ± 0.003	0.913 ± 0.005	0.873 ± 0.014	0.873 ± 0.005	0.915 ± 0.002	0.869 ± 0.014	0.888 ± 0.003	0.895 ± 0.006
	Audio+Image	0.909 ± 0.003	0.897 ± 0.003	0.919 ± 0.001	0.841 ± 0.003	0.866 ± 0.006	0.916 ± 0.001	0.873 ± 0.004	0.889 ± 0.002
	Text+Audio+Image	0.903 ± 0.008	0.904 ± 0.009	0.907 ± 0.003	0.888 ± 0.003	0.896 ± 0.002	0.903 ± 0.004	0.905 ± 0.001	0.901 ± 0.003
MLP	Text	0.941 ± 0.002	0.922 ± 0.007	0.860 ± 0.017	0.854 ± 0.008	0.891 ± 0.004	0.852 ± 0.012	0.850 ± 0.002	0.881 ± 0.003
	Image	0.924 ± 0.003	0.926 ± 0.007	0.885 ± 0.012	0.823 ± 0.007	0.881 ± 0.006	0.854 ± 0.010	0.847 ± 0.007	0.877 ± 0.005
	Audio	0.850 ± 0.023	0.895 ± 0.006	0.932 ± 0.002	0.819 ± 0.008	0.842 ± 0.006	0.907 ± 0.003	0.841 ± 0.003	0.869 ± 0.004
	Text+Image	0.909 ± 0.006	0.906 ± 0.008	0.867 ± 0.008	0.900 ± 0.006	0.890 ± 0.006	0.875 ± 0.006	0.881 ± 0.005	0.890 ± 0.004
	Text+Audio	0.934 ± 0.003	0.918 ± 0.004	0.854 ± 0.024	0.878 ± 0.005	0.915 ± 0.004	0.872 ± 0.017	0.883 ± 0.004	0.893 ± 0.007
	Audio+Image	0.912 ± 0.004	0.902 ± 0.007	0.923 ± 0.003	0.853 ± 0.007	0.885 ± 0.007	0.920 ± 0.003	0.882 ± 0.005	0.897 ± 0.003
	Text+Audio+Image	0.901 ± 0.006	0.910 ± 0.006	0.912 ± 0.003	0.888 ± 0.002	0.893 ± 0.007	0.905 ± 0.005	0.907 ± 0.005	0.902 ± 0.002
Phi-4-multimodal									
Linear	Text	0.940 ± 0.002	0.883 ± 0.009	0.637 ± 0.034	0.707 ± 0.006	0.834 ± 0.002	0.605 ± 0.011	0.701 ± 0.006	0.758 ± 0.007
	Image	0.885 ± 0.016	0.922 ± 0.002	0.752 ± 0.024	0.753 ± 0.003	0.826 ± 0.013	0.726 ± 0.006	0.754 ± 0.002	0.803 ± 0.006
	Audio	0.908 ± 0.003	0.876 ± 0.004	0.935 ± 0.001	0.717 ± 0.007	0.834 ± 0.006	0.719 ± 0.015	0.727 ± 0.007	0.817 ± 0.005
	Text+Image	0.858 ± 0.005	0.879 ± 0.002	0.619 ± 0.010	0.854 ± 0.001	0.795 ± 0.010	0.743 ± 0.006	0.849 ± 0.001	0.799 ± 0.003
	Text+Audio	0.929 ± 0.002	0.898 ± 0.009	0.841 ± 0.015	0.762 ± 0.006	0.899 ± 0.003	0.716 ± 0.005	0.765 ± 0.007	0.830 ± 0.005
	Audio+Image	0.860 ± 0.005	0.883 ± 0.005	0.907 ± 0.001	0.816 ± 0.004	0.804 ± 0.005	0.899 ± 0.002	0.824 ± 0.003	0.856 ± 0.002
	Text+Audio+Image	0.869 ± 0.006	0.887 ± 0.001	0.845 ± 0.013	0.852 ± 0.001	0.801 ± 0.014	0.852 ± 0.003	0.858 ± 0.001	0.852 ± 0.005
MLP	Text	0.944 ± 0.002	0.874 ± 0.023	0.780 ± 0.044	0.709 ± 0.012	0.844 ± 0.009	0.618 ± 0.034	0.708 ± 0.011	0.782 ± 0.015
	Image	0.909 ± 0.008	0.924 ± 0.003	0.836 ± 0.007	0.724 ± 0.007	0.842 ± 0.010	0.736 ± 0.013	0.729 ± 0.007	0.814 ± 0.004
	Audio	0.895 ± 0.003	0.875 ± 0.020	0.935 ± 0.003	0.734 ± 0.010	0.841 ± 0.014	0.732 ± 0.040	0.746 ± 0.010	0.823 ± 0.010
	Text+Image	0.848 ± 0.014	0.849 ± 0.014	0.649 ± 0.046	0.850 ± 0.003	0.817 ± 0.009	0.746 ± 0.019	0.849 ± 0.003	0.801 ± 0.012
	Text+Audio	0.929 ± 0.005	0.879 ± 0.015	0.845 ± 0.024	0.759 ± 0.013	0.896 ± 0.003	0.682 ± 0.014	0.761 ± 0.013	0.821 ± 0.008
	Audio+Image	0.861 ± 0.005	0.857 ± 0.008	0.913 ± 0.002	0.797 ± 0.004	0.810 ± 0.018	0.898 ± 0.002	0.811 ± 0.005	0.849 ± 0.003
	Text+Audio+Image	0.854 ± 0.010	0.860 ± 0.022	0.843 ± 0.011	0.849 ± 0.004	0.823 ± 0.012	0.849 ± 0.017	0.856 ± 0.003	0.848 ± 0.007

Table 13: **Cross-configuration evaluation results on Aegis-Omni (last layer).** The table shows AUC for linear and MLP probes across five omni-models. Entries report the mean ± standard deviation across five seeds, rounded to three decimal places. Rows indicate the training modality and columns indicate the evaluation modality. The final column reports the average over all seven evaluation modalities.

Monitor	Train ↓ / Eval →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
Qwen2.5-Omni-3B									
Linear	Text	0.901 ± 0.001	0.908 ± 0.002	0.522 ± 0.014	0.745 ± 0.007	0.728 ± 0.017	0.687 ± 0.013	0.738 ± 0.009	0.747 ± 0.007
	Image	0.841 ± 0.014	0.915 ± 0.001	0.658 ± 0.043	0.734 ± 0.010	0.783 ± 0.015	0.753 ± 0.017	0.743 ± 0.010	0.775 ± 0.011
	Audio	0.834 ± 0.004	0.854 ± 0.005	0.892 ± 0.001	0.714 ± 0.011	0.783 ± 0.001	0.816 ± 0.010	0.743 ± 0.010	0.805 ± 0.004
	Text+Image	0.860 ± 0.003	0.876 ± 0.002	0.509 ± 0.025	0.829 ± 0.003	0.710 ± 0.018	0.753 ± 0.005	0.810 ± 0.003	0.764 ± 0.006
	Text+Audio	0.884 ± 0.002	0.886 ± 0.005	0.844 ± 0.007	0.764 ± 0.010	0.868 ± 0.002	0.783 ± 0.020	0.774 ± 0.012	0.829 ± 0.007
	Audio+Image	0.857 ± 0.005	0.891 ± 0.003	0.850 ± 0.009	0.791 ± 0.007	0.806 ± 0.008	0.880 ± 0.002	0.806 ± 0.005	0.840 ± 0.004
	Text+Audio+Image	0.859 ± 0.002	0.878 ± 0.005	0.852 ± 0.006	0.827 ± 0.004	0.825 ± 0.002	0.835 ± 0.007	0.841 ± 0.003	0.845 ± 0.001
MLP	Text	0.903 ± 0.003	0.905 ± 0.004	0.536 ± 0.053	0.752 ± 0.015	0.721 ± 0.024	0.729 ± 0.013	0.746 ± 0.015	0.756 ± 0.009
	Image	0.841 ± 0.014	0.916 ± 0.002	0.699 ± 0.025	0.755 ± 0.021	0.801 ± 0.010	0.765 ± 0.004	0.763 ± 0.019	0.792 ± 0.010
	Audio	0.847 ± 0.013	0.861 ± 0.006	0.894 ± 0.002	0.735 ± 0.022	0.798 ± 0.006	0.825 ± 0.012	0.755 ± 0.023	0.816 ± 0.009
	Text+Image	0.844 ± 0.003	0.874 ± 0.010	0.576 ± 0.045	0.834 ± 0.002	0.726 ± 0.018	0.781 ± 0.011	0.814 ± 0.004	0.778 ± 0.010
	Text+Audio	0.886 ± 0.003	0.871 ± 0.018	0.846 ± 0.014	0.769 ± 0.024	0.872 ± 0.003	0.805 ± 0.039	0.777 ± 0.029	0.832 ± 0.015
	Audio+Image	0.850 ± 0.005	0.892 ± 0.004	0.851 ± 0.010	0.792 ± 0.002	0.809 ± 0.006	0.887 ± 0.001	0.809 ± 0.001	0.841 ± 0.001
	Text+Audio+Image	0.844 ± 0.005	0.879 ± 0.004	0.865 ± 0.003	0.831 ± 0.004	0.813 ± 0.004	0.856 ± 0.005	0.847 ± 0.002	0.848 ± 0.002
Qwen2.5-Omni-7B									
Linear	Text	0.905 ± 0.002	0.908 ± 0.002	0.672 ± 0.022	0.751 ± 0.014	0.777 ± 0.012	0.772 ± 0.012	0.760 ± 0.012	0.792 ± 0.006
	Image	0.882 ± 0.005	0.913 ± 0.003	0.729 ± 0.014	0.755 ± 0.006	0.791 ± 0.007	0.782 ± 0.013	0.761 ± 0.005	0.802 ± 0.005
	Audio	0.848 ± 0.012	0.873 ± 0.006	0.890 ± 0.002	0.730 ± 0.013	0.796 ± 0.007	0.845 ± 0.003	0.756 ± 0.011	0.820 ± 0.006
	Text+Image	0.876 ± 0.002	0.882 ± 0.005	0.641 ± 0.020	0.837 ± 0.002	0.763 ± 0.006	0.777 ± 0.008	0.828 ± 0.003	0.801 ± 0.003
	Text+Audio	0.888 ± 0.004	0.889 ± 0.002	0.848 ± 0.010	0.799 ± 0.006	0.873 ± 0.004	0.834 ± 0.005	0.810 ± 0.005	0.849 ± 0.003
	Audio+Image	0.866 ± 0.007	0.882 ± 0.005	0.876 ± 0.005	0.809 ± 0.005	0.823 ± 0.004	0.882 ± 0.003	0.821 ± 0.004	0.851 ± 0.003
	Text+Audio+Image	0.867 ± 0.003	0.887 ± 0.004	0.872 ± 0.005	0.835 ± 0.002	0.836 ± 0.001	0.848 ± 0.001	0.849 ± 0.002	0.856 ± 0.001
MLP	Text	0.906 ± 0.002	0.901 ± 0.004	0.687 ± 0.044	0.736 ± 0.011	0.791 ± 0.023	0.781 ± 0.013	0.748 ± 0.012	0.793 ± 0.013
	Image	0.869 ± 0.011	0.914 ± 0.001	0.719 ± 0.027	0.749 ± 0.011	0.782 ± 0.017	0.790 ± 0.015	0.756 ± 0.011	0.797 ± 0.010
	Audio	0.858 ± 0.011	0.869 ± 0.008	0.891 ± 0.002	0.744 ± 0.007	0.809 ± 0.004	0.847 ± 0.004	0.767 ± 0.007	0.827 ± 0.003
	Text+Image	0.874 ± 0.003	0.889 ± 0.006	0.642 ± 0.040	0.839 ± 0.002	0.765 ± 0.024	0.781 ± 0.010	0.829 ± 0.004	0.802 ± 0.010
	Text+Audio	0.886 ± 0.004	0.886 ± 0.004	0.848 ± 0.011	0.791 ± 0.008	0.877 ± 0.003	0.827 ± 0.004	0.802 ± 0.005	0.845 ± 0.003
	Audio+Image	0.855 ± 0.014	0.877 ± 0.003	0.852 ± 0.024	0.798 ± 0.010	0.793 ± 0.020	0.887 ± 0.003	0.811 ± 0.007	0.839 ± 0.005
	Text+Audio+Image	0.862 ± 0.006	0.893 ± 0.005	0.871 ± 0.005	0.837 ± 0.004	0.830 ± 0.005	0.848 ± 0.006	0.850 ± 0.003	0.856 ± 0.002
Qwen3-Omni-30B									
Linear	Text	0.909 ± 0.002	0.792 ± 0.016	0.726 ± 0.012	0.780 ± 0.024	0.849 ± 0.013	0.681 ± 0.023	0.757 ± 0.024	0.785 ± 0.012
	Image	0.885 ± 0.003	0.905 ± 0.002	0.767 ± 0.017	0.819 ± 0.003	0.857 ± 0.004	0.774 ± 0.009	0.814 ± 0.002	0.832 ± 0.004
	Audio	0.842 ± 0.006	0.774 ± 0.024	0.890 ± 0.002	0.683 ± 0.016	0.819 ± 0.005	0.766 ± 0.028	0.688 ± 0.023	0.780 ± 0.012
	Text+Image	0.882 ± 0.003	0.874 ± 0.004	0.717 ± 0.018	0.869 ± 0.001	0.856 ± 0.006	0.805 ± 0.012	0.849 ± 0.003	0.836 ± 0.005
	Text+Audio	0.894 ± 0.004	0.831 ± 0.017	0.844 ± 0.005	0.806 ± 0.011	0.884 ± 0.001	0.768 ± 0.022	0.805 ± 0.009	0.833 ± 0.009
	Audio+Image	0.856 ± 0.005	0.861 ± 0.011	0.870 ± 0.006	0.831 ± 0.005	0.840 ± 0.006	0.871 ± 0.001	0.837 ± 0.004	0.852 ± 0.003
	Text+Audio+Image	0.883 ± 0.002	0.876 ± 0.006	0.851 ± 0.001	0.863 ± 0.002	0.869 ± 0.002	0.844 ± 0.003	0.861 ± 0.001	0.864 ± 0.001
MLP	Text	0.908 ± 0.003	0.801 ± 0.029	0.707 ± 0.045	0.799 ± 0.009	0.821 ± 0.023	0.697 ± 0.023	0.761 ± 0.017	0.785 ± 0.019
	Image	0.872 ± 0.006	0.906 ± 0.002	0.738 ± 0.025	0.805 ± 0.003	0.847 ± 0.006	0.761 ± 0.015	0.801 ± 0.003	0.819 ± 0.005
	Audio	0.866 ± 0.002	0.754 ± 0.024	0.893 ± 0.003	0.694 ± 0.008	0.831 ± 0.004	0.776 ± 0.028	0.700 ± 0.015	0.788 ± 0.009
	Text+Image	0.865 ± 0.012	0.871 ± 0.007	0.697 ± 0.012	0.873 ± 0.004	0.827 ± 0.008	0.809 ± 0.015	0.847 ± 0.003	0.827 ± 0.004
	Text+Audio	0.897 ± 0.009	0.818 ± 0.020	0.852 ± 0.014	0.809 ± 0.014	0.887 ± 0.002	0.802 ± 0.010	0.806 ± 0.010	0.839 ± 0.006
	Audio+Image	0.861 ± 0.007	0.859 ± 0.018	0.826 ± 0.060	0.817 ± 0.004	0.818 ± 0.028	0.873 ± 0.003	0.826 ± 0.004	0.840 ± 0.013
	Text+Audio+Image	0.869 ± 0.004	0.866 ± 0.011	0.839 ± 0.017	0.864 ± 0.005	0.850 ± 0.005	0.841 ± 0.004	0.861 ± 0.005	0.856 ± 0.006
Phi-4-multimodal									
Linear	Text	0.904 ± 0.002	0.850 ± 0.006	0.626 ± 0.039	0.688 ± 0.005	0.760 ± 0.014	0.678 ± 0.008	0.693 ± 0.005	0.743 ± 0.005
	Image	0.840 ± 0.017	0.893 ± 0.001	0.725 ± 0.058	0.728 ± 0.005	0.775 ± 0.015	0.739 ± 0.004	0.737 ± 0.004	0.777 ± 0.014
	Audio	0.818 ± 0.014	0.774 ± 0.015	0.892 ± 0.002	0.679 ± 0.008	0.750 ± 0.015	0.725 ± 0.017	0.697 ± 0.012	0.762 ± 0.007
	Text+Image	0.788 ± 0.008	0.855 ± 0.005	0.700 ± 0.033	0.810 ± 0.005	0.754 ± 0.018	0.806 ± 0.006	0.813 ± 0.005	0.789 ± 0.009
	Text+Audio	0.865 ± 0.004	0.859 ± 0.003	0.841 ± 0.011	0.761 ± 0.012	0.851 ± 0.003	0.765 ± 0.015	0.764 ± 0.012	0.815 ± 0.006
	Audio+Image	0.663 ± 0.025	0.835 ± 0.011	0.833 ± 0.016	0.785 ± 0.011	0.689 ± 0.015	0.853 ± 0.005	0.804 ± 0.009	0.780 ± 0.012
	Text+Audio+Image	0.771 ± 0.018	0.858 ± 0.005	0.858 ± 0.007	0.809 ± 0.005	0.771 ± 0.008	0.837 ± 0.004	0.822 ± 0.004	0.818 ± 0.006
MLP	Text	0.903 ± 0.004	0.839 ± 0.011	0.659 ± 0.044	0.686 ± 0.011	0.756 ± 0.022	0.690 ± 0.012	0.694 ± 0.011	0.747 ± 0.008
	Image	0.804 ± 0.029	0.897 ± 0.001	0.662 ± 0.089	0.734 ± 0.009	0.755 ± 0.042	0.740 ± 0.007	0.740 ± 0.007	0.762 ± 0.022
	Audio	0.811 ± 0.018	0.779 ± 0.036	0.893 ± 0.003	0.704 ± 0.033	0.763 ± 0.021	0.746 ± 0.031	0.721 ± 0.031	0.774 ± 0.020
	Text+Image	0.724 ± 0.022	0.846 ± 0.014	0.711 ± 0.042	0.815 ± 0.005	0.733 ± 0.017	0.804 ± 0.007	0.815 ± 0.005	0.778 ± 0.012
	Text+Audio	0.857 ± 0.004	0.860 ± 0.006	0.832 ± 0.016	0.755 ± 0.020	0.851 ± 0.003	0.748 ± 0.024	0.758 ± 0.021	0.809 ± 0.009
	Audio+Image	0.676 ± 0.047	0.833 ± 0.014	0.810 ± 0.034	0.780 ± 0.014	0.703 ± 0.041	0.854 ± 0.007	0.799 ± 0.011	0.779 ± 0.018
	Text+Audio+Image	0.726 ± 0.032	0.850 ± 0.009	0.855 ± 0.015	0.812 ± 0.009	0.758 ± 0.017	0.836 ± 0.007	0.824 ± 0.006	0.808 ± 0.007
Interactive-Omni-8B									
Linear	Text	0.910 ± 0.001	0.907 ± 0.002	0.728 ± 0.015	0.790 ± 0.015	0.800 ± 0.004	0.806 ± 0.006	0.804 ± 0.007	0.821 ± 0.005
	Image	0.888 ± 0.005	0.915 ± 0.002	0.771 ± 0.021	0.785 ± 0.012	0.826 ± 0.008	0.815 ± 0.007	0.807 ± 0.008	0.830 ± 0.007
	Audio	0.821 ± 0.008	0.853 ± 0.006	0.892 ± 0.002	0.737 ± 0.024	0.785 ± 0.004	0.843 ± 0.007	0.787 ± 0.012	0.817 ± 0.007
	Text+Image	0.881 ± 0.005	0.871 ± 0.009	0.703 ± 0.022	0.860 ± 0.002	0.819 ± 0.013	0.822 ± 0.007	0.844 ± 0.004	0.829 ± 0.006
	Text+Audio	0.898 ± 0.002	0.868 ± 0.010	0.845 ± 0.007	0.831 ± 0.012	0.874 ± 0.004	0.848 ± 0.004	0.840 ± 0.008	0.858 ± 0.004
	Audio+Image	0.866 ± 0.006	0.865 ± 0.011	0.871 ± 0.005	0.801 ± 0.007	0.804 ± 0.004	0.875 ± 0.001	0.827 ± 0.004	0.844 ± 0.003
	Text+Audio+Image	0.877 ± 0.004	0.882 ± 0.006	0.866 ± 0.005	0.837 ± 0.006	0.848 ± 0.003	0.859 ± 0.003	0.854 ± 0.003	0.860 ± 0.002
MLP	Text	0.911 ± 0.004	0.905 ± 0.002	0.739 ± 0.023	0.776 ± 0.020	0.809 ± 0.008	0.821 ± 0.017	0.806 ± 0.009	0.824 ± 0.007
	Image	0.893 ± 0.005	0.913 ± 0.005	0.747 ± 0.026	0.785 ± 0.012	0.820 ± 0.009	0.820 ± 0.010	0.806 ± 0.003	0.826 ± 0.006
	Audio	0.850 ± 0.011	0.859 ± 0.005	0.893 ± 0.008	0.765 ± 0.007	0.803 ± 0.005	0.845 ± 0.006	0.796 ± 0.003	0.830 ± 0.004
	Text+Image	0.880 ± 0.003	0.879 ± 0.014	0.690 ± 0.034	0.864 ± 0.002	0.796 ± 0.018	0.821 ± 0.006	0.836 ± 0.006	0.824 ± 0.010
	Text+Audio	0.897 ± 0.009	0.879 ± 0.007	0.849 ± 0.016	0.841 ± 0.004	0.882 ± 0.003	0.854 ± 0.006	0.845 ± 0.003	0.864 ± 0.005
	Audio+Image	0.864 ± 0.010	0.874 ± 0.007	0.865 ± 0.013	0.796 ± 0.008	0.799 ± 0.010	0.881 ± 0.001	0.825 ± 0.005	0.844 ± 0.004
	Text+Audio+Image	0.871 ± 0.009	0.889 ± 0.007	0.871 ± 0.005	0.838 ± 0.004	0.841 ± 0.005	0.864 ± 0.004	0.860 ± 0.001	0.862 ± 0.002

B.2 Additional Metrics

We additionally report accuracy and F1 score in Tables 14–17. Unlike AUC, both metrics depend on the classification threshold, for which we use the standard value of 0.5. Nevertheless, they largely reproduce the qualitative patterns observed with AUC. Linear and MLP probes remain broadly comparable, with neither consistently outperforming the other. Transfer also remains asymmetric both across singleton modalities and between singleton and combined configurations.

Table 14: **Cross-configuration evaluation result on Aegis-Omni (Accuracy).**

Monitor	Train ↓ / Eval →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
Qwen2.5-Omni-3B									
Linear	Text	0.826 ± 0.003	0.830 ± 0.007	0.468 ± 0.002	0.631 ± 0.040	0.597 ± 0.025	0.561 ± 0.032	0.614 ± 0.039	0.647 ± 0.019
	Image	0.570 ± 0.097	0.837 ± 0.003	0.489 ± 0.022	0.584 ± 0.014	0.564 ± 0.052	0.609 ± 0.015	0.590 ± 0.011	0.606 ± 0.025
	Audio	0.524 ± 0.018	0.728 ± 0.035	0.828 ± 0.004	0.515 ± 0.034	0.619 ± 0.007	0.604 ± 0.066	0.526 ± 0.036	0.621 ± 0.026
	Text+Image	0.728 ± 0.066	0.795 ± 0.005	0.468 ± 0.020	0.754 ± 0.005	0.648 ± 0.015	0.538 ± 0.005	0.713 ± 0.009	0.663 ± 0.010
	Text+Audio	0.799 ± 0.004	0.769 ± 0.028	0.782 ± 0.008	0.671 ± 0.020	0.794 ± 0.001	0.699 ± 0.019	0.682 ± 0.024	0.742 ± 0.011
	Audio+Image	0.765 ± 0.017	0.807 ± 0.007	0.748 ± 0.041	0.635 ± 0.024	0.726 ± 0.021	0.802 ± 0.003	0.647 ± 0.024	0.733 ± 0.014
	Text+Audio+Image	0.616 ± 0.008	0.797 ± 0.006	0.588 ± 0.020	0.755 ± 0.004	0.642 ± 0.021	0.755 ± 0.009	0.764 ± 0.002	0.703 ± 0.007
MLP	Text	0.829 ± 0.008	0.810 ± 0.028	0.466 ± 0.002	0.613 ± 0.058	0.579 ± 0.026	0.545 ± 0.049	0.597 ± 0.059	0.634 ± 0.031
	Image	0.575 ± 0.036	0.838 ± 0.003	0.480 ± 0.011	0.585 ± 0.010	0.552 ± 0.025	0.601 ± 0.007	0.588 ± 0.007	0.603 ± 0.010
	Audio	0.549 ± 0.049	0.732 ± 0.018	0.829 ± 0.003	0.532 ± 0.018	0.633 ± 0.010	0.592 ± 0.033	0.544 ± 0.021	0.630 ± 0.012
	Text+Image	0.658 ± 0.086	0.792 ± 0.009	0.466 ± 0.002	0.757 ± 0.002	0.570 ± 0.042	0.570 ± 0.018	0.719 ± 0.006	0.647 ± 0.021
	Text+Audio	0.792 ± 0.008	0.757 ± 0.046	0.777 ± 0.005	0.691 ± 0.034	0.798 ± 0.004	0.724 ± 0.052	0.698 ± 0.034	0.748 ± 0.023
	Audio+Image	0.695 ± 0.042	0.803 ± 0.004	0.770 ± 0.034	0.641 ± 0.008	0.706 ± 0.020	0.810 ± 0.002	0.662 ± 0.013	0.727 ± 0.009
	Text+Audio+Image	0.692 ± 0.063	0.795 ± 0.002	0.622 ± 0.055	0.756 ± 0.005	0.696 ± 0.065	0.769 ± 0.009	0.767 ± 0.004	0.728 ± 0.027
Qwen2.5-Omni-7B									
Linear	Text	0.831 ± 0.004	0.827 ± 0.009	0.511 ± 0.038	0.534 ± 0.009	0.667 ± 0.013	0.600 ± 0.012	0.541 ± 0.007	0.644 ± 0.009
	Image	0.610 ± 0.048	0.842 ± 0.002	0.490 ± 0.010	0.578 ± 0.013	0.580 ± 0.022	0.621 ± 0.013	0.581 ± 0.014	0.615 ± 0.013
	Audio	0.614 ± 0.077	0.783 ± 0.015	0.825 ± 0.002	0.507 ± 0.009	0.670 ± 0.029	0.684 ± 0.022	0.522 ± 0.010	0.658 ± 0.008
	Text+Image	0.778 ± 0.019	0.791 ± 0.007	0.474 ± 0.004	0.761 ± 0.005	0.681 ± 0.013	0.629 ± 0.021	0.745 ± 0.006	0.694 ± 0.007
	Text+Audio	0.784 ± 0.027	0.811 ± 0.004	0.776 ± 0.011	0.624 ± 0.024	0.802 ± 0.001	0.757 ± 0.007	0.647 ± 0.026	0.743 ± 0.005
	Audio+Image	0.765 ± 0.038	0.786 ± 0.008	0.786 ± 0.009	0.715 ± 0.017	0.746 ± 0.019	0.807 ± 0.006	0.728 ± 0.010	0.762 ± 0.005
	Text+Audio+Image	0.753 ± 0.045	0.796 ± 0.004	0.763 ± 0.020	0.759 ± 0.004	0.766 ± 0.005	0.745 ± 0.011	0.771 ± 0.002	0.765 ± 0.006
MLP	Text	0.831 ± 0.006	0.819 ± 0.014	0.487 ± 0.018	0.535 ± 0.016	0.621 ± 0.023	0.569 ± 0.021	0.538 ± 0.015	0.629 ± 0.012
	Image	0.607 ± 0.064	0.843 ± 0.006	0.488 ± 0.010	0.571 ± 0.017	0.566 ± 0.005	0.613 ± 0.004	0.576 ± 0.017	0.609 ± 0.011
	Audio	0.621 ± 0.081	0.766 ± 0.032	0.825 ± 0.005	0.512 ± 0.019	0.697 ± 0.010	0.676 ± 0.041	0.522 ± 0.021	0.660 ± 0.023
	Text+Image	0.717 ± 0.069	0.799 ± 0.003	0.520 ± 0.071	0.757 ± 0.005	0.669 ± 0.014	0.650 ± 0.029	0.739 ± 0.004	0.693 ± 0.009
	Text+Audio	0.777 ± 0.034	0.803 ± 0.015	0.777 ± 0.009	0.636 ± 0.027	0.798 ± 0.003	0.751 ± 0.007	0.650 ± 0.027	0.742 ± 0.009
	Audio+Image	0.713 ± 0.067	0.779 ± 0.017	0.777 ± 0.024	0.697 ± 0.029	0.699 ± 0.031	0.809 ± 0.002	0.717 ± 0.022	0.742 ± 0.015
	Text+Audio+Image	0.704 ± 0.096	0.799 ± 0.006	0.756 ± 0.039	0.750 ± 0.011	0.736 ± 0.036	0.732 ± 0.019	0.766 ± 0.006	0.749 ± 0.013
Qwen3-Omni-30B									
Linear	Text	0.835 ± 0.002	0.640 ± 0.064	0.672 ± 0.015	0.606 ± 0.052	0.768 ± 0.011	0.569 ± 0.043	0.603 ± 0.052	0.670 ± 0.028
	Image	0.557 ± 0.073	0.837 ± 0.004	0.575 ± 0.041	0.590 ± 0.029	0.603 ± 0.053	0.640 ± 0.020	0.594 ± 0.025	0.628 ± 0.031
	Audio	0.578 ± 0.053	0.603 ± 0.038	0.828 ± 0.004	0.591 ± 0.036	0.686 ± 0.016	0.592 ± 0.034	0.597 ± 0.037	0.639 ± 0.019
	Text+Image	0.748 ± 0.035	0.792 ± 0.009	0.541 ± 0.034	0.797 ± 0.003	0.692 ± 0.034	0.733 ± 0.019	0.725 ± 0.007	0.718 ± 0.015
	Text+Audio	0.817 ± 0.005	0.738 ± 0.009	0.788 ± 0.007	0.686 ± 0.034	0.809 ± 0.002	0.629 ± 0.032	0.698 ± 0.024	0.738 ± 0.011
	Audio+Image	0.635 ± 0.097	0.758 ± 0.021	0.785 ± 0.048	0.726 ± 0.033	0.705 ± 0.044	0.796 ± 0.003	0.727 ± 0.026	0.733 ± 0.013
	Text+Audio+Image	0.785 ± 0.028	0.795 ± 0.017	0.734 ± 0.049	0.791 ± 0.007	0.776 ± 0.041	0.706 ± 0.020	0.789 ± 0.004	0.768 ± 0.016
MLP	Text	0.833 ± 0.004	0.619 ± 0.114	0.572 ± 0.051	0.619 ± 0.094	0.723 ± 0.008	0.572 ± 0.076	0.599 ± 0.088	0.648 ± 0.057
	Image	0.527 ± 0.066	0.836 ± 0.005	0.530 ± 0.044	0.557 ± 0.026	0.562 ± 0.078	0.588 ± 0.025	0.560 ± 0.025	0.594 ± 0.023
	Audio	0.573 ± 0.088	0.604 ± 0.039	0.828 ± 0.004	0.592 ± 0.030	0.672 ± 0.052	0.590 ± 0.025	0.598 ± 0.034	0.637 ± 0.021
	Text+Image	0.560 ± 0.069	0.785 ± 0.008	0.551 ± 0.079	0.794 ± 0.005	0.590 ± 0.093	0.699 ± 0.034	0.717 ± 0.011	0.671 ± 0.041
	Text+Audio	0.807 ± 0.021	0.645 ± 0.090	0.779 ± 0.031	0.608 ± 0.121	0.807 ± 0.008	0.594 ± 0.087	0.619 ± 0.122	0.694 ± 0.063
	Audio+Image	0.559 ± 0.111	0.759 ± 0.029	0.566 ± 0.032	0.727 ± 0.028	0.584 ± 0.082	0.794 ± 0.007	0.725 ± 0.021	0.673 ± 0.029
	Text+Audio+Image	0.605 ± 0.096	0.772 ± 0.012	0.749 ± 0.110	0.789 ± 0.005	0.667 ± 0.093	0.720 ± 0.030	0.785 ± 0.003	0.727 ± 0.043
Phi-4-multimodal									
Linear	Text	0.830 ± 0.005	0.709 ± 0.021	0.507 ± 0.048	0.476 ± 0.003	0.642 ± 0.027	0.472 ± 0.002	0.475 ± 0.004	0.587 ± 0.010
	Image	0.699 ± 0.024	0.830 ± 0.003	0.658 ± 0.029	0.544 ± 0.010	0.686 ± 0.027	0.544 ± 0.007	0.547 ± 0.009	0.644 ± 0.007
	Audio	0.662 ± 0.051	0.546 ± 0.028	0.827 ± 0.004	0.462 ± 0.001	0.594 ± 0.013	0.463 ± 0.001	0.463 ± 0.001	0.574 ± 0.009
	Text+Image	0.689 ± 0.011	0.771 ± 0.009	0.616 ± 0.070	0.736 ± 0.006	0.668 ± 0.023	0.740 ± 0.008	0.739 ± 0.006	0.708 ± 0.014
	Text+Audio	0.779 ± 0.017	0.734 ± 0.026	0.756 ± 0.017	0.551 ± 0.050	0.779 ± 0.003	0.553 ± 0.048	0.551 ± 0.048	0.672 ± 0.023
	Audio+Image	0.466 ± 0.015	0.752 ± 0.013	0.721 ± 0.038	0.656 ± 0.013	0.545 ± 0.035	0.782 ± 0.008	0.687 ± 0.008	0.658 ± 0.015
	Text+Audio+Image	0.641 ± 0.070	0.773 ± 0.017	0.736 ± 0.035	0.728 ± 0.009	0.678 ± 0.035	0.754 ± 0.006	0.743 ± 0.007	0.722 ± 0.011
MLP	Text	0.828 ± 0.004	0.691 ± 0.065	0.536 ± 0.065	0.470 ± 0.010	0.629 ± 0.039	0.468 ± 0.008	0.469 ± 0.010	0.584 ± 0.013
	Image	0.698 ± 0.054	0.830 ± 0.007	0.589 ± 0.094	0.546 ± 0.025	0.601 ± 0.052	0.543 ± 0.022	0.549 ± 0.026	0.622 ± 0.034
	Audio	0.551 ± 0.060	0.491 ± 0.016	0.825 ± 0.002	0.461 ± 0.000	0.581 ± 0.006	0.461 ± 0.000	0.461 ± 0.000	0.547 ± 0.009
	Text+Image	0.593 ± 0.048	0.760 ± 0.017	0.601 ± 0.041	0.739 ± 0.006	0.613 ± 0.058	0.732 ± 0.007	0.740 ± 0.006	0.682 ± 0.022
	Text+Audio	0.767 ± 0.012	0.703 ± 0.038	0.754 ± 0.015	0.492 ± 0.026	0.779 ± 0.006	0.488 ± 0.022	0.491 ± 0.026	0.639 ± 0.015
	Audio+Image	0.493 ± 0.053	0.748 ± 0.021	0.683 ± 0.077	0.656 ± 0.017	0.545 ± 0.063	0.781 ± 0.010	0.689 ± 0.011	0.656 ± 0.023
	Text+Audio+Image	0.586 ± 0.098	0.759 ± 0.010	0.740 ± 0.054	0.730 ± 0.006	0.618 ± 0.074	0.754 ± 0.008	0.746 ± 0.005	0.705 ± 0.028
Interactive-Omni-8B									
Linear	Text	0.834 ± 0.003	0.802 ± 0.018	0.552 ± 0.063	0.593 ± 0.028	0.637 ± 0.035	0.545 ± 0.034	0.576 ± 0.033	0.648 ± 0.028
	Image	0.532 ± 0.024	0.835 ± 0.004	0.527 ± 0.024	0.605 ± 0.005	0.544 ± 0.029	0.603 ± 0.028	0.630 ± 0.013	0.611 ± 0.015
	Audio	0.517 ± 0.013	0.698 ± 0.037	0.830 ± 0.004	0.524 ± 0.027	0.652 ± 0.012	0.729 ± 0.017	0.636 ± 0.023	0.655 ± 0.015
	Text+Image	0.700 ± 0.064	0.793 ± 0.007	0.557 ± 0.018	0.782 ± 0.004	0.709 ± 0.023	0.610 ± 0.028	0.633 ± 0.030	0.683 ± 0.018
	Text+Audio	0.811 ± 0.006	0.778 ± 0.015	0.785 ± 0.006	0.722 ± 0.026	0.799 ± 0.004	0.751 ± 0.022	0.763 ± 0.012	0.773 ± 0.011
	Audio+Image	0.585 ± 0.100	0.792 ± 0.016	0.796 ± 0.015	0.685 ± 0.016	0.681 ± 0.059	0.805 ± 0.005	0.739 ± 0.002	0.726 ± 0.022
	Text+Audio+Image	0.740 ± 0.081	0.802 ± 0.010	0.784 ± 0.025	0.757 ± 0.009	0.752 ± 0.036	0.778 ± 0.003	0.774 ± 0.003	0.770 ± 0.015
MLP	Text	0.829 ± 0.008	0.768 ± 0.036	0.501 ± 0.016	0.563 ± 0.028	0.603 ± 0.025	0.521 ± 0.029	0.552 ± 0.034	0.620 ± 0.023
	Image	0.623 ± 0.095	0.836 ± 0.006	0.533 ± 0.019	0.585 ± 0.020	0.573 ± 0.054	0.572 ± 0.028	0.612 ± 0.017	0.619 ± 0.029
	Audio	0.554 ± 0.089	0.718 ± 0.032	0.826 ± 0.010	0.550 ± 0.016	0.667 ± 0.042	0.719 ± 0.038	0.647 ± 0.023	0.669 ± 0.013
	Text+Image	0.670 ± 0.043	0.791 ± 0.012	0.557 ± 0.038	0.780 ± 0.006	0.691 ± 0.033	0.624 ± 0.026	0.667 ± 0.019	0.683 ± 0.014
	Text+Audio	0.800 ± 0.030	0.786 ± 0.019	0.787 ± 0.005	0.749 ± 0.015	0.803 ± 0.001	0.750 ± 0.014	0.763 ± 0.013	0.777 ± 0.010
	Audio+Image	0.662 ± 0.136	0.801 ± 0.012	0.788 ± 0.025	0.656 ± 0.043	0.687 ± 0.052	0.805 ± 0.004	0.728 ± 0.010	0.733 ± 0.032
	Text+Audio+Image	0.700 ± 0.095	0.810 ± 0.008	0.795 ± 0.011	0.755 ± 0.005	0.729 ± 0.028	0.782 ± 0.010	0.779 ± 0.003	0.764 ± 0.015

Table 15: **Cross-configuration evaluation result on Aegis-Omni (F1).**

Monitor	Train ↓ / Eval →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
Qwen2.5-Omni-3B									
	Text	0.836 ± 0.003	0.843 ± 0.015	0.070 ± 0.033	0.535 ± 0.115	0.435 ± 0.059	0.349 ± 0.102	0.487 ± 0.115	0.508 ± 0.056
	Image	0.332 ± 0.230	0.848 ± 0.003	0.175 ± 0.109	0.399 ± 0.039	0.351 ± 0.146	0.493 ± 0.036	0.413 ± 0.030	0.430 ± 0.068
	Audio	0.219 ± 0.055	0.688 ± 0.064	0.843 ± 0.004	0.193 ± 0.115	0.508 ± 0.016	0.433 ± 0.168	0.227 ± 0.113	0.444 ± 0.070
Linear	**Text+Image**	0.783 ± 0.030	0.810 ± 0.008	0.149 ± 0.141	0.777 ± 0.006	0.605 ± 0.065	0.292 ± 0.015	0.702 ± 0.014	0.589 ± 0.028
	Text+Audio	0.796 ± 0.007	0.812 ± 0.015	0.819 ± 0.005	0.703 ± 0.043	0.813 ± 0.002	0.726 ± 0.034	0.705 ± 0.048	0.768 ± 0.016
	Audio+Image	0.770 ± 0.039	0.827 ± 0.004	0.795 ± 0.019	0.534 ± 0.064	0.766 ± 0.011	0.818 ± 0.003	0.557 ± 0.063	0.724 ± 0.022
	Text+Audio+Image	0.734 ± 0.004	0.816 ± 0.009	0.722 ± 0.009	0.778 ± 0.005	0.745 ± 0.009	0.778 ± 0.018	0.787 ± 0.002	0.766 ± 0.005
	Text	0.839 ± 0.009	0.812 ± 0.047	0.043 ± 0.026	0.478 ± 0.159	0.385 ± 0.076	0.286 ± 0.154	0.434 ± 0.166	0.468 ± 0.087
	Image	0.372 ± 0.102	0.848 ± 0.002	0.114 ± 0.056	0.402 ± 0.030	0.319 ± 0.083	0.466 ± 0.021	0.409 ± 0.023	0.419 ± 0.035
	Audio	0.286 ± 0.133	0.703 ± 0.036	0.842 ± 0.002	0.245 ± 0.059	0.538 ± 0.021	0.407 ± 0.090	0.280 ± 0.064	0.471 ± 0.032
MLP	**Text+Image**	0.560 ± 0.191	0.809 ± 0.012	0.041 ± 0.029	0.781 ± 0.005	0.373 ± 0.129	0.379 ± 0.050	0.719 ± 0.012	0.523 ± 0.056
	Text+Audio	0.792 ± 0.023	0.805 ± 0.026	0.816 ± 0.006	0.717 ± 0.035	0.819 ± 0.004	0.774 ± 0.021	0.724 ± 0.022	0.778 ± 0.010
	Audio+Image	0.637 ± 0.086	0.826 ± 0.001	0.796 ± 0.020	0.551 ± 0.026	0.695 ± 0.058	0.827 ± 0.003	0.591 ± 0.031	0.703 ± 0.024
	Text+Audio+Image	0.728 ± 0.029	0.815 ± 0.007	0.739 ± 0.026	0.776 ± 0.008	0.764 ± 0.025	0.805 ± 0.003	0.790 ± 0.004	0.774 ± 0.008
Qwen2.5-Omni-7B									
	Text	0.842 ± 0.005	0.829 ± 0.014	0.252 ± 0.106	0.248 ± 0.028	0.601 ± 0.024	0.440 ± 0.033	0.268 ± 0.023	0.497 ± 0.024
	Image	0.445 ± 0.101	0.852 ± 0.003	0.150 ± 0.033	0.377 ± 0.036	0.393 ± 0.063	0.491 ± 0.030	0.386 ± 0.040	0.442 ± 0.032
	Audio	0.451 ± 0.184	0.778 ± 0.025	0.841 ± 0.002	0.166 ± 0.029	0.617 ± 0.060	0.622 ± 0.043	0.211 ± 0.031	0.526 ± 0.020
Linear	**Text+Image**	0.807 ± 0.009	0.806 ± 0.011	0.217 ± 0.052	0.780 ± 0.004	0.659 ± 0.038	0.548 ± 0.051	0.755 ± 0.005	0.653 ± 0.023
	Text+Audio	0.770 ± 0.041	0.826 ± 0.006	0.813 ± 0.006	0.500 ± 0.064	0.820 ± 0.002	0.779 ± 0.009	0.549 ± 0.061	0.723 ± 0.015
	Audio+Image	0.779 ± 0.072	0.814 ± 0.004	0.818 ± 0.006	0.714 ± 0.036	0.765 ± 0.046	0.821 ± 0.006	0.733 ± 0.023	0.777 ± 0.015
	Text+Audio+Image	0.730 ± 0.081	0.813 ± 0.009	0.811 ± 0.012	0.771 ± 0.004	0.782 ± 0.015	0.794 ± 0.004	0.789 ± 0.002	0.784 ± 0.013
	Text	0.842 ± 0.008	0.824 ± 0.021	0.183 ± 0.064	0.250 ± 0.049	0.498 ± 0.051	0.355 ± 0.063	0.260 ± 0.048	0.459 ± 0.035
	Image	0.439 ± 0.133	0.852 ± 0.007	0.143 ± 0.042	0.356 ± 0.051	0.356 ± 0.014	0.478 ± 0.015	0.371 ± 0.050	0.428 ± 0.024
	Audio	0.462 ± 0.171	0.752 ± 0.054	0.840 ± 0.007	0.180 ± 0.065	0.663 ± 0.019	0.603 ± 0.085	0.211 ± 0.070	0.530 ± 0.055
MLP	**Text+Image**	0.770 ± 0.022	0.816 ± 0.009	0.328 ± 0.216	0.773 ± 0.003	0.658 ± 0.064	0.596 ± 0.062	0.748 ± 0.007	0.670 ± 0.048
	Text+Audio	0.763 ± 0.053	0.829 ± 0.007	0.813 ± 0.009	0.528 ± 0.065	0.817 ± 0.007	0.769 ± 0.013	0.558 ± 0.061	0.725 ± 0.021
	Audio+Image	0.760 ± 0.026	0.811 ± 0.010	0.806 ± 0.018	0.678 ± 0.066	0.723 ± 0.028	0.822 ± 0.004	0.708 ± 0.051	0.758 ± 0.015
	Text+Audio+Image	0.654 ± 0.207	0.819 ± 0.006	0.807 ± 0.019	0.760 ± 0.013	0.740 ± 0.075	0.787 ± 0.007	0.782 ± 0.007	0.764 ± 0.038
Qwen3-Omni-30B									
	Text	0.846 ± 0.003	0.729 ± 0.015	0.672 ± 0.060	0.727 ± 0.019	0.786 ± 0.016	0.702 ± 0.007	0.724 ± 0.014	0.741 ± 0.007
	Image	0.294 ± 0.197	0.848 ± 0.004	0.421 ± 0.133	0.405 ± 0.078	0.426 ± 0.132	0.562 ± 0.058	0.411 ± 0.066	0.481 ± 0.088
	Audio	0.378 ± 0.135	0.726 ± 0.013	0.844 ± 0.004	0.711 ± 0.010	0.641 ± 0.031	0.723 ± 0.013	0.711 ± 0.013	0.676 ± 0.021
Linear	**Text+Image**	0.716 ± 0.061	0.808 ± 0.010	0.337 ± 0.130	0.812 ± 0.004	0.627 ± 0.069	0.765 ± 0.009	0.697 ± 0.012	0.680 ± 0.036
	Text+Audio	0.824 ± 0.008	0.757 ± 0.035	0.811 ± 0.014	0.747 ± 0.021	0.825 ± 0.003	0.734 ± 0.012	0.749 ± 0.029	0.778 ± 0.012
	Audio+Image	0.505 ± 0.239	0.742 ± 0.036	0.811 ± 0.021	0.714 ± 0.068	0.690 ± 0.092	0.812 ± 0.003	0.711 ± 0.051	0.712 ± 0.034
	Text+Audio+Image	0.801 ± 0.022	0.816 ± 0.013	0.796 ± 0.026	0.809 ± 0.009	0.813 ± 0.021	0.777 ± 0.010	0.806 ± 0.006	0.802 ± 0.009
	Text	0.843 ± 0.005	0.460 ± 0.304	0.412 ± 0.139	0.502 ± 0.286	0.711 ± 0.019	0.435 ± 0.306	0.449 ± 0.275	0.545 ± 0.178
	Image	0.212 ± 0.186	0.846 ± 0.006	0.301 ± 0.175	0.316 ± 0.077	0.311 ± 0.209	0.427 ± 0.078	0.321 ± 0.070	0.391 ± 0.074
	Audio	0.334 ± 0.223	0.723 ± 0.014	0.843 ± 0.005	0.719 ± 0.010	0.600 ± 0.104	0.722 ± 0.010	0.718 ± 0.008	0.666 ± 0.045
MLP	**Text+Image**	0.309 ± 0.193	0.789 ± 0.013	0.341 ± 0.292	0.808 ± 0.008	0.386 ± 0.255	0.698 ± 0.092	0.685 ± 0.021	0.574 ± 0.123
	Text+Audio	0.808 ± 0.034	0.563 ± 0.259	0.803 ± 0.054	0.491 ± 0.360	0.821 ± 0.012	0.524 ± 0.328	0.505 ± 0.343	0.645 ± 0.187
	Audio+Image	0.289 ± 0.270	0.751 ± 0.051	0.356 ± 0.080	0.727 ± 0.056	0.378 ± 0.210	0.810 ± 0.008	0.716 ± 0.044	0.575 ± 0.077
	Text+Audio+Image	0.415 ± 0.243	0.770 ± 0.021	0.724 ± 0.209	0.807 ± 0.010	0.567 ± 0.228	0.781 ± 0.015	0.803 ± 0.006	0.695 ± 0.095
Phi-4-multimodal									
	Text	0.842 ± 0.005	0.654 ± 0.038	0.305 ± 0.120	0.056 ± 0.009	0.581 ± 0.073	0.042 ± 0.009	0.055 ± 0.013	0.362 ± 0.030
	Image	0.751 ± 0.034	0.841 ± 0.003	0.686 ± 0.049	0.295 ± 0.028	0.678 ± 0.077	0.294 ± 0.021	0.300 ± 0.025	0.549 ± 0.018
	Audio	0.587 ± 0.123	0.292 ± 0.080	0.843 ± 0.004	0.005 ± 0.004	0.451 ± 0.029	0.009 ± 0.005	0.007 ± 0.004	0.313 ± 0.020
Linear	**Text+Image**	0.732 ± 0.042	0.798 ± 0.009	0.604 ± 0.187	0.758 ± 0.006	0.687 ± 0.067	0.769 ± 0.009	0.762 ± 0.006	0.730 ± 0.044
	Text+Audio	0.790 ± 0.032	0.704 ± 0.051	0.804 ± 0.009	0.305 ± 0.152	0.800 ± 0.002	0.309 ± 0.145	0.304 ± 0.147	0.574 ± 0.071
	Audio+Image	0.051 ± 0.076	0.757 ± 0.022	0.683 ± 0.072	0.601 ± 0.027	0.322 ± 0.116	0.799 ± 0.008	0.655 ± 0.015	0.552 ± 0.041
	Text+Audio+Image	0.614 ± 0.224	0.797 ± 0.012	0.780 ± 0.020	0.743 ± 0.011	0.679 ± 0.105	0.795 ± 0.005	0.764 ± 0.008	0.739 ± 0.049
	Text	0.840 ± 0.005	0.625 ± 0.137	0.376 ± 0.158	0.035 ± 0.041	0.556 ± 0.085	0.027 ± 0.033	0.032 ± 0.039	0.356 ± 0.035
	Image	0.699 ± 0.126	0.842 ± 0.007	0.498 ± 0.245	0.298 ± 0.081	0.461 ± 0.126	0.289 ± 0.074	0.308 ± 0.084	0.485 ± 0.091
	Audio	0.300 ± 0.176	0.107 ± 0.054	0.838 ± 0.003	0.000 ± 0.001	0.418 ± 0.011	0.000 ± 0.001	0.000 ± 0.001	0.238 ± 0.028
MLP	**Text+Image**	0.472 ± 0.166	0.775 ± 0.025	0.526 ± 0.129	0.761 ± 0.007	0.523 ± 0.171	0.757 ± 0.011	0.762 ± 0.010	0.654 ± 0.065
	Text+Audio	0.764 ± 0.022	0.642 ± 0.073	0.801 ± 0.005	0.111 ± 0.091	0.799 ± 0.006	0.098 ± 0.080	0.110 ± 0.090	0.475 ± 0.048
	Audio+Image	0.136 ± 0.199	0.752 ± 0.037	0.615 ± 0.175	0.595 ± 0.041	0.305 ± 0.202	0.797 ± 0.009	0.655 ± 0.023	0.551 ± 0.067
	Text+Audio+Image	0.419 ± 0.293	0.775 ± 0.030	0.749 ± 0.106	0.738 ± 0.007	0.502 ± 0.207	0.793 ± 0.005	0.764 ± 0.006	0.677 ± 0.087
Interactive-Omni-8B									
	Text	0.844 ± 0.004	0.793 ± 0.028	0.330 ± 0.152	0.410 ± 0.077	0.526 ± 0.076	0.279 ± 0.098	0.363 ± 0.086	0.506 ± 0.070
	Image	0.241 ± 0.071	0.845 ± 0.004	0.256 ± 0.086	0.445 ± 0.012	0.280 ± 0.086	0.441 ± 0.072	0.509 ± 0.032	0.431 ± 0.042
	Audio	0.213 ± 0.039	0.638 ± 0.069	0.845 ± 0.003	0.226 ± 0.085	0.585 ± 0.024	0.707 ± 0.028	0.530 ± 0.051	0.535 ± 0.034
Linear	**Text+Image**	0.776 ± 0.033	0.815 ± 0.011	0.445 ± 0.061	0.801 ± 0.005	0.679 ± 0.053	0.468 ± 0.068	0.511 ± 0.066	0.642 ± 0.033
	Text+Audio	0.811 ± 0.014	0.806 ± 0.012	0.817 ± 0.003	0.779 ± 0.009	0.816 ± 0.003	0.797 ± 0.009	0.790 ± 0.008	0.802 ± 0.006
	Audio+Image	0.357 ± 0.257	0.805 ± 0.021	0.823 ± 0.007	0.649 ± 0.037	0.626 ± 0.117	0.822 ± 0.004	0.731 ± 0.004	0.688 ± 0.054
	Text+Audio+Image	0.695 ± 0.152	0.823 ± 0.008	0.822 ± 0.014	0.783 ± 0.012	0.748 ± 0.064	0.812 ± 0.002	0.788 ± 0.004	0.782 ± 0.029
	Text	0.839 ± 0.006	0.741 ± 0.059	0.191 ± 0.069	0.327 ± 0.081	0.442 ± 0.065	0.208 ± 0.098	0.297 ± 0.102	0.435 ± 0.064
	Image	0.448 ± 0.237	0.846 ± 0.007	0.293 ± 0.071	0.393 ± 0.056	0.355 ± 0.159	0.363 ± 0.073	0.459 ± 0.044	0.451 ± 0.079
	Audio	0.288 ± 0.223	0.676 ± 0.060	0.841 ± 0.010	0.308 ± 0.045	0.606 ± 0.086	0.689 ± 0.075	0.556 ± 0.057	0.566 ± 0.033
MLP	**Text+Image**	0.761 ± 0.021	0.815 ± 0.021	0.447 ± 0.133	0.797 ± 0.007	0.663 ± 0.076	0.500 ± 0.063	0.589 ± 0.039	0.653 ± 0.043
	Text+Audio	0.795 ± 0.048	0.822 ± 0.009	0.823 ± 0.005	0.786 ± 0.016	0.819 ± 0.003	0.800 ± 0.007	0.793 ± 0.006	0.805 ± 0.009
	Audio+Image	0.525 ± 0.312	0.817 ± 0.012	0.813 ± 0.019	0.581 ± 0.104	0.658 ± 0.118	0.821 ± 0.004	0.712 ± 0.019	0.704 ± 0.079
	Text+Audio+Image	0.624 ± 0.184	0.830 ± 0.008	0.828 ± 0.006	0.777 ± 0.009	0.713 ± 0.051	0.816 ± 0.006	0.795 ± 0.003	0.769 ± 0.033

Table 16: **Cross-configuration evaluation result on WGM-Omni (Accuracy).**

Monitor	Train ↓ / Eval →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
Qwen2.5-Omni-3B									
Linear	Text	0.861 ± 0.003	0.850 ± 0.019	0.580 ± 0.011	0.703 ± 0.017	0.810 ± 0.007	0.650 ± 0.020	0.699 ± 0.024	0.736 ± 0.006
	Image	0.731 ± 0.022	0.873 ± 0.003	0.576 ± 0.008	0.699 ± 0.010	0.764 ± 0.022	0.671 ± 0.016	0.716 ± 0.011	0.719 ± 0.011
	Audio	0.741 ± 0.005	0.811 ± 0.014	0.868 ± 0.003	0.683 ± 0.034	0.742 ± 0.009	0.762 ± 0.064	0.699 ± 0.032	0.758 ± 0.021
	Text+Image	0.726 ± 0.009	0.848 ± 0.004	0.574 ± 0.003	0.799 ± 0.002	0.722 ± 0.006	0.664 ± 0.008	0.794 ± 0.002	0.732 ± 0.004
	Text+Audio	0.834 ± 0.005	0.805 ± 0.019	0.765 ± 0.009	0.569 ± 0.039	0.840 ± 0.005	0.682 ± 0.034	0.574 ± 0.038	0.724 ± 0.019
	Audio+Image	0.771 ± 0.009	0.792 ± 0.019	0.851 ± 0.004	0.750 ± 0.005	0.762 ± 0.009	0.844 ± 0.003	0.756 ± 0.007	0.789 ± 0.004
	Text+Audio+Image	0.729 ± 0.021	0.853 ± 0.003	0.817 ± 0.003	0.789 ± 0.005	0.746 ± 0.021	0.781 ± 0.010	0.812 ± 0.002	0.790 ± 0.006
MLP	Text	0.866 ± 0.003	0.842 ± 0.024	0.596 ± 0.015	0.697 ± 0.021	0.813 ± 0.002	0.648 ± 0.020	0.693 ± 0.016	0.736 ± 0.010
	Image	0.796 ± 0.032	0.877 ± 0.004	0.680 ± 0.058	0.686 ± 0.004	0.790 ± 0.016	0.698 ± 0.019	0.703 ± 0.006	0.747 ± 0.018
	Audio	0.745 ± 0.017	0.806 ± 0.020	0.867 ± 0.005	0.712 ± 0.017	0.748 ± 0.016	0.804 ± 0.019	0.719 ± 0.019	0.771 ± 0.012
	Text+Image	0.700 ± 0.026	0.842 ± 0.005	0.572 ± 0.004	0.799 ± 0.004	0.688 ± 0.023	0.687 ± 0.011	0.795 ± 0.007	0.726 ± 0.007
	Text+Audio	0.863 ± 0.013	0.815 ± 0.025	0.708 ± 0.050	0.616 ± 0.051	0.846 ± 0.008	0.659 ± 0.069	0.613 ± 0.055	0.731 ± 0.031
	Audio+Image	0.769 ± 0.033	0.796 ± 0.024	0.817 ± 0.064	0.733 ± 0.024	0.751 ± 0.026	0.844 ± 0.007	0.747 ± 0.017	0.779 ± 0.014
	Text+Audio+Image	0.715 ± 0.037	0.839 ± 0.004	0.815 ± 0.011	0.788 ± 0.005	0.696 ± 0.042	0.813 ± 0.008	0.805 ± 0.006	0.782 ± 0.013
Qwen2.5-Omni-7B									
Linear	Text	0.873 ± 0.008	0.873 ± 0.011	0.713 ± 0.039	0.744 ± 0.009	0.740 ± 0.005	0.654 ± 0.025	0.739 ± 0.016	0.762 ± 0.014
	Image	0.656 ± 0.014	0.887 ± 0.003	0.742 ± 0.009	0.654 ± 0.010	0.706 ± 0.020	0.690 ± 0.004	0.678 ± 0.013	0.716 ± 0.007
	Audio	0.664 ± 0.018	0.848 ± 0.006	0.800 ± 0.006	0.676 ± 0.015	0.711 ± 0.015	0.823 ± 0.008	0.699 ± 0.016	0.746 ± 0.010
	Text+Image	0.766 ± 0.004	0.865 ± 0.004	0.811 ± 0.005	0.810 ± 0.005	0.829 ± 0.006	0.766 ± 0.015	0.804 ± 0.006	0.807 ± 0.001
	Text+Audio	0.840 ± 0.009	0.869 ± 0.004	0.798 ± 0.015	0.693 ± 0.010	0.852 ± 0.001	0.747 ± 0.023	0.727 ± 0.014	0.790 ± 0.007
	Audio+Image	0.710 ± 0.015	0.874 ± 0.004	0.802 ± 0.007	0.728 ± 0.008	0.744 ± 0.011	0.854 ± 0.002	0.749 ± 0.003	0.780 ± 0.005
	Text+Audio+Image	0.732 ± 0.009	0.863 ± 0.003	0.806 ± 0.006	0.809 ± 0.005	0.806 ± 0.004	0.803 ± 0.017	0.816 ± 0.005	0.805 ± 0.003
MLP	Text	0.877 ± 0.006	0.869 ± 0.009	0.774 ± 0.043	0.701 ± 0.019	0.794 ± 0.019	0.718 ± 0.025	0.715 ± 0.019	0.778 ± 0.008
	Image	0.718 ± 0.020	0.886 ± 0.001	0.771 ± 0.008	0.658 ± 0.031	0.770 ± 0.009	0.686 ± 0.015	0.676 ± 0.033	0.738 ± 0.010
	Audio	0.699 ± 0.021	0.837 ± 0.016	0.801 ± 0.013	0.697 ± 0.008	0.751 ± 0.008	0.831 ± 0.005	0.715 ± 0.008	0.762 ± 0.008
	Text+Image	0.751 ± 0.007	0.857 ± 0.008	0.812 ± 0.009	0.811 ± 0.005	0.806 ± 0.009	0.768 ± 0.015	0.808 ± 0.007	0.802 ± 0.004
	Text+Audio	0.855 ± 0.007	0.876 ± 0.003	0.775 ± 0.021	0.781 ± 0.007	0.853 ± 0.006	0.751 ± 0.017	0.796 ± 0.014	0.812 ± 0.006
	Audio+Image	0.753 ± 0.033	0.877 ± 0.005	0.808 ± 0.007	0.727 ± 0.017	0.781 ± 0.015	0.859 ± 0.004	0.749 ± 0.013	0.793 ± 0.010
	Text+Audio+Image	0.746 ± 0.008	0.855 ± 0.009	0.808 ± 0.004	0.802 ± 0.007	0.801 ± 0.009	0.790 ± 0.018	0.819 ± 0.003	0.803 ± 0.003
Qwen3-Omni-30B									
Linear	Text	0.875 ± 0.003	0.727 ± 0.053	0.592 ± 0.021	0.656 ± 0.083	0.848 ± 0.007	0.588 ± 0.014	0.657 ± 0.068	0.706 ± 0.032
	Image	0.822 ± 0.027	0.854 ± 0.003	0.635 ± 0.023	0.697 ± 0.015	0.813 ± 0.024	0.673 ± 0.021	0.710 ± 0.019	0.744 ± 0.017
	Audio	0.774 ± 0.009	0.651 ± 0.007	0.868 ± 0.001	0.619 ± 0.022	0.802 ± 0.009	0.612 ± 0.012	0.629 ± 0.027	0.708 ± 0.008
	Text+Image	0.854 ± 0.010	0.841 ± 0.009	0.615 ± 0.019	0.845 ± 0.003	0.857 ± 0.001	0.606 ± 0.040	0.821 ± 0.002	0.777 ± 0.009
	Text+Audio	0.866 ± 0.005	0.752 ± 0.052	0.774 ± 0.015	0.708 ± 0.069	0.862 ± 0.002	0.631 ± 0.035	0.709 ± 0.070	0.757 ± 0.027
	Audio+Image	0.797 ± 0.046	0.821 ± 0.015	0.792 ± 0.049	0.774 ± 0.007	0.609 ± 0.064	0.845 ± 0.004	0.781 ± 0.004	0.774 ± 0.022
	Text+Audio+Image	0.841 ± 0.020	0.836 ± 0.010	0.819 ± 0.011	0.819 ± 0.004	0.841 ± 0.014	0.768 ± 0.015	0.829 ± 0.003	0.822 ± 0.004
MLP	Text	0.877 ± 0.006	0.666 ± 0.057	0.612 ± 0.027	0.643 ± 0.050	0.847 ± 0.009	0.565 ± 0.034	0.632 ± 0.046	0.692 ± 0.016
	Image	0.792 ± 0.099	0.866 ± 0.006	0.667 ± 0.093	0.715 ± 0.025	0.707 ± 0.118	0.687 ± 0.012	0.739 ± 0.021	0.739 ± 0.039
	Audio	0.778 ± 0.020	0.701 ± 0.026	0.874 ± 0.004	0.581 ± 0.021	0.803 ± 0.016	0.577 ± 0.009	0.578 ± 0.019	0.699 ± 0.005
	Text+Image	0.741 ± 0.020	0.828 ± 0.014	0.525 ± 0.026	0.844 ± 0.006	0.760 ± 0.026	0.619 ± 0.023	0.820 ± 0.003	0.734 ± 0.010
	Text+Audio	0.881 ± 0.010	0.743 ± 0.084	0.744 ± 0.025	0.650 ± 0.098	0.864 ± 0.007	0.563 ± 0.088	0.643 ± 0.107	0.727 ± 0.037
	Audio+Image	0.755 ± 0.065	0.819 ± 0.026	0.843 ± 0.015	0.766 ± 0.008	0.747 ± 0.030	0.847 ± 0.003	0.771 ± 0.016	0.793 ± 0.005
	Text+Audio+Image	0.716 ± 0.044	0.805 ± 0.033	0.699 ± 0.074	0.820 ± 0.011	0.746 ± 0.043	0.739 ± 0.032	0.831 ± 0.006	0.765 ± 0.025
Phi-4-multimodal									
Linear	Text	0.875 ± 0.003	0.762 ± 0.033	0.586 ± 0.004	0.574 ± 0.009	0.762 ± 0.016	0.569 ± 0.005	0.571 ± 0.007	0.671 ± 0.008
	Image	0.743 ± 0.030	0.851 ± 0.005	0.604 ± 0.031	0.697 ± 0.007	0.710 ± 0.029	0.668 ± 0.007	0.693 ± 0.008	0.709 ± 0.009
	Audio	0.765 ± 0.042	0.708 ± 0.035	0.869 ± 0.002	0.566 ± 0.008	0.690 ± 0.015	0.579 ± 0.012	0.568 ± 0.009	0.678 ± 0.008
	Text+Image	0.618 ± 0.017	0.795 ± 0.005	0.561 ± 0.002	0.791 ± 0.003	0.717 ± 0.017	0.685 ± 0.006	0.769 ± 0.005	0.705 ± 0.005
	Text+Audio	0.864 ± 0.012	0.817 ± 0.007	0.776 ± 0.017	0.614 ± 0.015	0.831 ± 0.005	0.617 ± 0.014	0.618 ± 0.014	0.734 ± 0.007
	Audio+Image	0.602 ± 0.019	0.735 ± 0.032	0.807 ± 0.012	0.711 ± 0.005	0.607 ± 0.020	0.836 ± 0.003	0.731 ± 0.003	0.718 ± 0.011
	Text+Audio+Image	0.594 ± 0.014	0.758 ± 0.010	0.692 ± 0.029	0.771 ± 0.001	0.702 ± 0.023	0.768 ± 0.004	0.794 ± 0.004	0.726 ± 0.010
MLP	Text	0.879 ± 0.003	0.736 ± 0.047	0.622 ± 0.039	0.566 ± 0.010	0.758 ± 0.015	0.564 ± 0.007	0.567 ± 0.011	0.670 ± 0.014
	Image	0.810 ± 0.044	0.857 ± 0.004	0.655 ± 0.060	0.648 ± 0.022	0.751 ± 0.030	0.657 ± 0.026	0.652 ± 0.021	0.718 ± 0.022
	Audio	0.638 ± 0.049	0.664 ± 0.050	0.870 ± 0.005	0.562 ± 0.011	0.692 ± 0.014	0.573 ± 0.027	0.564 ± 0.015	0.652 ± 0.013
	Text+Image	0.742 ± 0.043	0.745 ± 0.032	0.607 ± 0.033	0.780 ± 0.008	0.722 ± 0.040	0.672 ± 0.018	0.752 ± 0.013	0.717 ± 0.017
	Text+Audio	0.859 ± 0.013	0.746 ± 0.035	0.775 ± 0.020	0.570 ± 0.022	0.828 ± 0.006	0.572 ± 0.020	0.572 ± 0.023	0.703 ± 0.015
	Audio+Image	0.728 ± 0.087	0.760 ± 0.023	0.826 ± 0.020	0.702 ± 0.014	0.718 ± 0.034	0.834 ± 0.005	0.725 ± 0.010	0.756 ± 0.017
	Text+Audio+Image	0.745 ± 0.040	0.759 ± 0.024	0.741 ± 0.054	0.777 ± 0.007	0.701 ± 0.085	0.756 ± 0.024	0.787 ± 0.004	0.752 ± 0.030
Interactive-Omni-8B									
Linear	Text	0.876 ± 0.002	0.797 ± 0.022	0.577 ± 0.013	0.617 ± 0.020	0.814 ± 0.008	0.578 ± 0.012	0.664 ± 0.032	0.703 ± 0.015
	Image	0.722 ± 0.018	0.868 ± 0.004	0.663 ± 0.033	0.724 ± 0.007	0.779 ± 0.016	0.738 ± 0.011	0.786 ± 0.005	0.754 ± 0.012
	Audio	0.727 ± 0.014	0.801 ± 0.014	0.868 ± 0.006	0.706 ± 0.008	0.774 ± 0.006	0.838 ± 0.008	0.753 ± 0.005	0.781 ± 0.006
	Text+Image	0.837 ± 0.014	0.846 ± 0.005	0.777 ± 0.023	0.842 ± 0.002	0.835 ± 0.004	0.802 ± 0.004	0.813 ± 0.006	0.822 ± 0.005
	Text+Audio	0.838 ± 0.005	0.845 ± 0.006	0.802 ± 0.031	0.759 ± 0.016	0.845 ± 0.008	0.792 ± 0.021	0.811 ± 0.012	0.813 ± 0.012
	Audio+Image	0.806 ± 0.038	0.788 ± 0.019	0.869 ± 0.002	0.731 ± 0.017	0.793 ± 0.007	0.851 ± 0.006	0.796 ± 0.004	0.805 ± 0.006
	Text+Audio+Image	0.771 ± 0.013	0.835 ± 0.010	0.845 ± 0.002	0.795 ± 0.012	0.816 ± 0.011	0.838 ± 0.006	0.836 ± 0.002	0.820 ± 0.004
MLP	Text	0.878 ± 0.001	0.820 ± 0.023	0.617 ± 0.039	0.687 ± 0.036	0.820 ± 0.011	0.618 ± 0.034	0.725 ± 0.034	0.738 ± 0.022
	Image	0.720 ± 0.008	0.865 ± 0.004	0.689 ± 0.044	0.706 ± 0.016	0.772 ± 0.005	0.731 ± 0.029	0.764 ± 0.004	0.750 ± 0.010
	Audio	0.754 ± 0.033	0.813 ± 0.014	0.872 ± 0.007	0.734 ± 0.016	0.751 ± 0.010	0.841 ± 0.005	0.765 ± 0.006	0.790 ± 0.004
	Text+Image	0.798 ± 0.022	0.834 ± 0.012	0.681 ± 0.064	0.835 ± 0.005	0.805 ± 0.017	0.744 ± 0.035	0.812 ± 0.005	0.787 ± 0.017
	Text+Audio	0.861 ± 0.014	0.845 ± 0.010	0.789 ± 0.017	0.796 ± 0.017	0.847 ± 0.011	0.753 ± 0.037	0.768 ± 0.043	0.808 ± 0.014
	Audio+Image	0.788 ± 0.063	0.808 ± 0.024	0.870 ± 0.004	0.747 ± 0.023	0.804 ± 0.019	0.856 ± 0.007	0.796 ± 0.011	0.810 ± 0.012
	Text+Audio+Image	0.771 ± 0.026	0.839 ± 0.007	0.851 ± 0.010	0.805 ± 0.007	0.807 ± 0.010	0.843 ± 0.008	0.839 ± 0.009	0.822 ± 0.007

Table 17: **Cross-configuration evaluation result on WGM-Omni (F1).**

Monitor	Train ↓ / Eval →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
Qwen2.5-Omni-3B									
Linear	Text	0.830 ± 0.003	0.812 ± 0.034	0.377 ± 0.054	0.656 ± 0.054	0.770 ± 0.014	0.494 ± 0.094	0.661 ± 0.044	0.657 ± 0.036
	Image	0.582 ± 0.053	0.847 ± 0.004	0.364 ± 0.049	0.549 ± 0.027	0.671 ± 0.054	0.560 ± 0.053	0.607 ± 0.030	0.597 ± 0.036
	Audio	0.666 ± 0.028	0.765 ± 0.029	0.845 ± 0.005	0.511 ± 0.128	0.680 ± 0.007	0.653 ± 0.145	0.554 ± 0.115	0.668 ± 0.057
	Text+Image	0.575 ± 0.022	0.815 ± 0.008	0.206 ± 0.039	0.769 ± 0.003	0.575 ± 0.017	0.471 ± 0.017	0.769 ± 0.005	0.597 ± 0.011
	Text+Audio	0.789 ± 0.007	0.807 ± 0.014	0.772 ± 0.006	0.666 ± 0.015	0.806 ± 0.007	0.724 ± 0.017	0.669 ± 0.015	0.747 ± 0.010
	Audio+Image	0.726 ± 0.024	0.786 ± 0.012	0.823 ± 0.005	0.696 ± 0.019	0.721 ± 0.023	0.817 ± 0.004	0.696 ± 0.022	0.752 ± 0.003
	Text+Audio+Image	0.579 ± 0.054	0.823 ± 0.004	0.793 ± 0.006	0.745 ± 0.009	0.640 ± 0.048	0.723 ± 0.021	0.784 ± 0.003	0.727 ± 0.016
MLP	Text	0.838 ± 0.005	0.806 ± 0.045	0.344 ± 0.077	0.638 ± 0.089	0.769 ± 0.008	0.485 ± 0.079	0.637 ± 0.083	0.645 ± 0.035
	Image	0.714 ± 0.063	0.853 ± 0.004	0.574 ± 0.128	0.538 ± 0.019	0.738 ± 0.044	0.615 ± 0.034	0.596 ± 0.019	0.661 ± 0.038
	Audio	0.723 ± 0.028	0.792 ± 0.013	0.843 ± 0.006	0.663 ± 0.022	0.731 ± 0.009	0.754 ± 0.041	0.677 ± 0.019	0.741 ± 0.010
	Text+Image	0.519 ± 0.079	0.811 ± 0.014	0.153 ± 0.037	0.770 ± 0.005	0.494 ± 0.071	0.518 ± 0.030	0.769 ± 0.008	0.576 ± 0.024
	Text+Audio	0.835 ± 0.020	0.813 ± 0.017	0.711 ± 0.056	0.681 ± 0.017	0.815 ± 0.010	0.705 ± 0.029	0.684 ± 0.022	0.749 ± 0.015
	Audio+Image	0.711 ± 0.059	0.785 ± 0.011	0.760 ± 0.126	0.651 ± 0.081	0.687 ± 0.059	0.818 ± 0.009	0.673 ± 0.055	0.726 ± 0.023
	Text+Audio+Image	0.558 ± 0.105	0.811 ± 0.009	0.769 ± 0.029	0.740 ± 0.005	0.528 ± 0.119	0.777 ± 0.018	0.777 ± 0.006	0.708 ± 0.038
Qwen2.5-Omni-7B									
Linear	Text	0.853 ± 0.009	0.859 ± 0.009	0.749 ± 0.023	0.712 ± 0.014	0.759 ± 0.004	0.697 ± 0.012	0.731 ± 0.011	0.766 ± 0.009
	Image	0.382 ± 0.044	0.866 ± 0.003	0.695 ± 0.011	0.398 ± 0.029	0.523 ± 0.050	0.607 ± 0.018	0.471 ± 0.034	0.563 ± 0.019
	Audio	0.397 ± 0.057	0.809 ± 0.011	0.725 ± 0.011	0.503 ± 0.056	0.557 ± 0.039	0.781 ± 0.014	0.562 ± 0.050	0.619 ± 0.031
	Text+Image	0.673 ± 0.013	0.836 ± 0.008	0.787 ± 0.009	0.780 ± 0.006	0.798 ± 0.010	0.753 ± 0.003	0.786 ± 0.007	0.773 ± 0.005
	Text+Audio	0.800 ± 0.014	0.841 ± 0.007	0.778 ± 0.017	0.489 ± 0.027	0.822 ± 0.002	0.678 ± 0.042	0.581 ± 0.034	0.713 ± 0.014
	Audio+Image	0.528 ± 0.040	0.851 ± 0.006	0.729 ± 0.014	0.611 ± 0.020	0.625 ± 0.024	0.830 ± 0.003	0.650 ± 0.012	0.689 ± 0.011
	Text+Audio+Image	0.593 ± 0.021	0.833 ± 0.006	0.768 ± 0.006	0.762 ± 0.007	0.751 ± 0.008	0.793 ± 0.010	0.783 ± 0.005	0.755 ± 0.003
MLP	Text	0.857 ± 0.009	0.845 ± 0.019	0.781 ± 0.024	0.554 ± 0.064	0.794 ± 0.013	0.684 ± 0.014	0.598 ± 0.060	0.730 ± 0.018
	Image	0.548 ± 0.050	0.865 ± 0.002	0.732 ± 0.014	0.414 ± 0.098	0.674 ± 0.019	0.587 ± 0.052	0.470 ± 0.102	0.613 ± 0.031
	Audio	0.496 ± 0.054	0.789 ± 0.029	0.723 ± 0.026	0.563 ± 0.028	0.649 ± 0.019	0.792 ± 0.006	0.608 ± 0.025	0.660 ± 0.017
	Text+Image	0.669 ± 0.022	0.826 ± 0.016	0.784 ± 0.016	0.784 ± 0.007	0.782 ± 0.013	0.752 ± 0.012	0.791 ± 0.007	0.770 ± 0.008
	Text+Audio	0.819 ± 0.012	0.855 ± 0.002	0.762 ± 0.025	0.728 ± 0.024	0.824 ± 0.007	0.759 ± 0.017	0.766 ± 0.016	0.788 ± 0.007
	Audio+Image	0.628 ± 0.072	0.855 ± 0.005	0.739 ± 0.013	0.630 ± 0.041	0.703 ± 0.032	0.835 ± 0.005	0.671 ± 0.032	0.723 ± 0.022
	Text+Audio+Image	0.643 ± 0.031	0.821 ± 0.016	0.774 ± 0.014	0.753 ± 0.012	0.762 ± 0.018	0.792 ± 0.013	0.791 ± 0.006	0.762 ± 0.009
Qwen3-Omni-30B									
Linear	Text	0.848 ± 0.004	0.559 ± 0.133	0.395 ± 0.022	0.358 ± 0.257	0.816 ± 0.010	0.275 ± 0.088	0.378 ± 0.226	0.518 ± 0.100
	Image	0.772 ± 0.049	0.821 ± 0.004	0.405 ± 0.120	0.509 ± 0.039	0.761 ± 0.051	0.503 ± 0.099	0.548 ± 0.050	0.617 ± 0.053
	Audio	0.720 ± 0.004	0.678 ± 0.008	0.845 ± 0.002	0.260 ± 0.087	0.771 ± 0.009	0.241 ± 0.044	0.292 ± 0.098	0.544 ± 0.033
	Text+Image	0.822 ± 0.016	0.817 ± 0.008	0.540 ± 0.043	0.814 ± 0.003	0.835 ± 0.004	0.660 ± 0.011	0.789 ± 0.004	0.754 ± 0.008
	Text+Audio	0.837 ± 0.007	0.621 ± 0.123	0.750 ± 0.040	0.626 ± 0.171	0.833 ± 0.003	0.525 ± 0.194	0.638 ± 0.161	0.690 ± 0.093
	Audio+Image	0.794 ± 0.027	0.799 ± 0.009	0.796 ± 0.033	0.721 ± 0.018	0.686 ± 0.032	0.820 ± 0.006	0.757 ± 0.007	0.767 ± 0.015
	Text+Audio+Image	0.800 ± 0.034	0.789 ± 0.017	0.816 ± 0.007	0.764 ± 0.009	0.827 ± 0.009	0.763 ± 0.005	0.796 ± 0.003	0.794 ± 0.007
MLP	Text	0.852 ± 0.008	0.390 ± 0.162	0.388 ± 0.112	0.342 ± 0.177	0.817 ± 0.014	0.221 ± 0.212	0.314 ± 0.181	0.475 ± 0.118
	Image	0.799 ± 0.061	0.838 ± 0.009	0.700 ± 0.042	0.574 ± 0.074	0.743 ± 0.068	0.650 ± 0.033	0.642 ± 0.064	0.706 ± 0.018
	Audio	0.735 ± 0.045	0.673 ± 0.010	0.851 ± 0.006	0.108 ± 0.088	0.781 ± 0.012	0.093 ± 0.044	0.096 ± 0.079	0.477 ± 0.035
	Text+Image	0.618 ± 0.046	0.797 ± 0.024	0.407 ± 0.087	0.814 ± 0.010	0.661 ± 0.052	0.655 ± 0.009	0.793 ± 0.004	0.678 ± 0.024
	Text+Audio	0.861 ± 0.012	0.590 ± 0.202	0.713 ± 0.056	0.575 ± 0.139	0.837 ± 0.009	0.486 ± 0.212	0.577 ± 0.133	0.663 ± 0.101
	Audio+Image	0.631 ± 0.152	0.788 ± 0.032	0.815 ± 0.021	0.724 ± 0.015	0.709 ± 0.054	0.819 ± 0.003	0.749 ± 0.010	0.748 ± 0.029
	Text+Audio+Image	0.547 ± 0.115	0.730 ± 0.064	0.501 ± 0.185	0.765 ± 0.018	0.624 ± 0.108	0.740 ± 0.006	0.798 ± 0.008	0.672 ± 0.063
Phi-4-multimodal									
Linear	Text	0.849 ± 0.004	0.652 ± 0.074	0.250 ± 0.026	0.093 ± 0.044	0.688 ± 0.040	0.067 ± 0.031	0.079 ± 0.038	0.383 ± 0.024
	Image	0.605 ± 0.071	0.817 ± 0.007	0.227 ± 0.121	0.567 ± 0.023	0.553 ± 0.074	0.484 ± 0.027	0.558 ± 0.025	0.545 ± 0.026
	Audio	0.653 ± 0.097	0.517 ± 0.095	0.846 ± 0.003	0.046 ± 0.036	0.509 ± 0.040	0.108 ± 0.058	0.055 ± 0.042	0.391 ± 0.026
	Text+Image	0.250 ± 0.059	0.739 ± 0.014	0.032 ± 0.012	0.761 ± 0.004	0.569 ± 0.046	0.667 ± 0.007	0.758 ± 0.004	0.539 ± 0.016
	Text+Audio	0.832 ± 0.022	0.768 ± 0.016	0.743 ± 0.023	0.261 ± 0.062	0.794 ± 0.006	0.288 ± 0.058	0.277 ± 0.059	0.566 ± 0.029
	Audio+Image	0.188 ± 0.071	0.588 ± 0.079	0.740 ± 0.021	0.567 ± 0.012	0.252 ± 0.076	0.806 ± 0.004	0.617 ± 0.007	0.537 ± 0.034
	Text+Audio+Image	0.158 ± 0.054	0.641 ± 0.022	0.488 ± 0.079	0.700 ± 0.003	0.543 ± 0.062	0.756 ± 0.005	0.761 ± 0.005	0.578 ± 0.029
MLP	Text	0.855 ± 0.005	0.588 ± 0.126	0.314 ± 0.133	0.048 ± 0.051	0.670 ± 0.030	0.048 ± 0.048	0.053 ± 0.056	0.368 ± 0.048
	Image	0.769 ± 0.092	0.827 ± 0.008	0.407 ± 0.203	0.461 ± 0.110	0.663 ± 0.076	0.464 ± 0.109	0.478 ± 0.107	0.581 ± 0.077
	Audio	0.304 ± 0.148	0.385 ± 0.151	0.847 ± 0.007	0.029 ± 0.055	0.511 ± 0.041	0.076 ± 0.119	0.038 ± 0.071	0.313 ± 0.049
	Text+Image	0.644 ± 0.125	0.732 ± 0.015	0.393 ± 0.176	0.754 ± 0.006	0.705 ± 0.024	0.674 ± 0.007	0.749 ± 0.008	0.664 ± 0.039
	Text+Audio	0.831 ± 0.025	0.612 ± 0.089	0.751 ± 0.037	0.067 ± 0.105	0.791 ± 0.012	0.081 ± 0.101	0.075 ± 0.109	0.458 ± 0.058
	Audio+Image	0.593 ± 0.294	0.725 ± 0.044	0.806 ± 0.036	0.568 ± 0.042	0.663 ± 0.128	0.809 ± 0.003	0.624 ± 0.026	0.684 ± 0.067
	Text+Audio+Image	0.683 ± 0.056	0.693 ± 0.036	0.701 ± 0.042	0.716 ± 0.012	0.698 ± 0.038	0.750 ± 0.016	0.760 ± 0.005	0.714 ± 0.018
Interactive-Omni-8B									
Linear	Text	0.851 ± 0.003	0.716 ± 0.041	0.094 ± 0.056	0.254 ± 0.072	0.772 ± 0.015	0.104 ± 0.052	0.403 ± 0.098	0.456 ± 0.046
	Image	0.554 ± 0.044	0.839 ± 0.005	0.434 ± 0.096	0.586 ± 0.018	0.689 ± 0.035	0.632 ± 0.027	0.730 ± 0.007	0.638 ± 0.031
	Audio	0.597 ± 0.038	0.784 ± 0.007	0.846 ± 0.007	0.595 ± 0.031	0.744 ± 0.007	0.813 ± 0.008	0.725 ± 0.006	0.729 ± 0.010
	Text+Image	0.793 ± 0.025	0.814 ± 0.005	0.700 ± 0.045	0.813 ± 0.002	0.812 ± 0.009	0.752 ± 0.014	0.791 ± 0.002	0.782 ± 0.011
	Text+Audio	0.793 ± 0.008	0.814 ± 0.009	0.774 ± 0.048	0.653 ± 0.039	0.814 ± 0.010	0.761 ± 0.035	0.763 ± 0.023	0.768 ± 0.022
	Audio+Image	0.739 ± 0.078	0.786 ± 0.011	0.852 ± 0.004	0.610 ± 0.042	0.770 ± 0.016	0.827 ± 0.008	0.751 ± 0.008	0.762 ± 0.018
	Text+Audio+Image	0.671 ± 0.029	0.799 ± 0.009	0.823 ± 0.005	0.729 ± 0.024	0.768 ± 0.023	0.820 ± 0.006	0.804 ± 0.004	0.773 ± 0.008
MLP	Text	0.853 ± 0.002	0.758 ± 0.043	0.257 ± 0.140	0.477 ± 0.103	0.787 ± 0.022	0.259 ± 0.119	0.573 ± 0.086	0.566 ± 0.065
	Image	0.549 ± 0.020	0.838 ± 0.004	0.488 ± 0.122	0.566 ± 0.056	0.674 ± 0.011	0.616 ± 0.073	0.705 ± 0.019	0.634 ± 0.030
	Audio	0.694 ± 0.077	0.795 ± 0.010	0.849 ± 0.009	0.642 ± 0.035	0.738 ± 0.019	0.815 ± 0.007	0.739 ± 0.009	0.753 ± 0.013
	Text+Image	0.730 ± 0.044	0.803 ± 0.013	0.469 ± 0.168	0.807 ± 0.006	0.754 ± 0.036	0.623 ± 0.079	0.770 ± 0.013	0.708 ± 0.042
	Text+Audio	0.829 ± 0.022	0.816 ± 0.019	0.776 ± 0.028	0.774 ± 0.010	0.817 ± 0.016	0.766 ± 0.020	0.769 ± 0.023	0.792 ± 0.009
	Audio+Image	0.699 ± 0.143	0.798 ± 0.015	0.851 ± 0.004	0.646 ± 0.056	0.770 ± 0.048	0.830 ± 0.010	0.755 ± 0.023	0.764 ± 0.033
	Text+Audio+Image	0.679 ± 0.063	0.804 ± 0.006	0.827 ± 0.016	0.752 ± 0.013	0.760 ± 0.026	0.823 ± 0.006	0.808 ± 0.011	0.779 ± 0.015

C Additional Results for three supervision regimes

C.1 Full Results for different $\mathcal{O}(n)$ and $\mathcal{O}(2^n)$ approaches

In this section, we provide the full AUC results for different approaches to $\mathcal{O}(n)$ and $\mathcal{O}(2^n)$ monitoring on WGM-Omni in Tables 18–22. In the main text, we report the $\mathcal{O}(n)$ - One Monitor Pooled Data result, as it performs best in most settings among the $\mathcal{O}(n)$ approaches. Similarly, for $\mathcal{O}(2^n)$ monitoring, we report $\mathcal{O}(2^n)$ - Configuration-Matched in the main text.

Table 18: **Qwen2.5-Omni-3B - $\mathcal{O}(n)$ and $\mathcal{O}(2^n)$:** Comparison of AUC using linear and MLP probes. The $\mathcal{O}(n)$ section uses Text, Image, and Audio training configurations; the $\mathcal{O}(2^n)$ section uses all seven modality combinations. Entries report mean ± standard deviation across five seeds, rounded to three decimals. The final column averages all seven test modalities.

Ensemble Method ↓ / Test →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
$\mathcal{O}(n)$: Text, Image, and Audio training								
Linear								
Avg. Prob.	0.909 ± 0.008	0.927 ± 0.004	0.827 ± 0.014	0.811 ± 0.006	0.881 ± 0.005	0.821 ± 0.023	0.816 ± 0.006	0.856 ± 0.006
Compositional Avg. Prob.	0.930 ± 0.002	0.933 ± 0.003	0.929 ± 0.002	0.805 ± 0.007	0.876 ± 0.006	0.846 ± 0.029	0.816 ± 0.006	0.876 ± 0.004
Max Prob.	0.911 ± 0.015	0.925 ± 0.005	0.901 ± 0.014	0.803 ± 0.009	0.875 ± 0.008	0.822 ± 0.036	0.806 ± 0.011	0.863 ± 0.008
Majority Vote	0.854 ± 0.005	0.887 ± 0.003	0.799 ± 0.018	0.756 ± 0.026	0.837 ± 0.011	0.763 ± 0.048	0.769 ± 0.018	0.809 ± 0.016
One Monitor Pooled Data	0.926 ± 0.001	0.931 ± 0.001	0.928 ± 0.001	0.809 ± 0.004	0.884 ± 0.001	0.893 ± 0.004	0.818 ± 0.004	0.884 ± 0.002
MLP								
Avg. Prob.	0.912 ± 0.006	0.928 ± 0.005	0.874 ± 0.014	0.811 ± 0.007	0.890 ± 0.007	0.848 ± 0.020	0.816 ± 0.006	0.868 ± 0.009
Compositional Avg. Prob.	0.935 ± 0.002	0.936 ± 0.003	0.933 ± 0.002	0.794 ± 0.007	0.879 ± 0.005	0.869 ± 0.016	0.816 ± 0.006	0.881 ± 0.005
Max Prob.	0.900 ± 0.013	0.926 ± 0.005	0.908 ± 0.024	0.806 ± 0.006	0.869 ± 0.016	0.851 ± 0.024	0.809 ± 0.007	0.867 ± 0.010
Majority Vote	0.872 ± 0.009	0.897 ± 0.007	0.844 ± 0.016	0.772 ± 0.010	0.856 ± 0.010	0.810 ± 0.023	0.779 ± 0.005	0.833 ± 0.008
One Monitor Pooled Data	0.935 ± 0.001	0.935 ± 0.003	0.931 ± 0.004	0.800 ± 0.005	0.890 ± 0.002	0.876 ± 0.010	0.809 ± 0.005	0.882 ± 0.003
$\mathcal{O}(2^n)$: All seven training configurations								
Linear								
Configuration-Matched	0.930 ± 0.002	0.933 ± 0.003	0.929 ± 0.002	0.864 ± 0.002	0.903 ± 0.001	0.909 ± 0.002	0.870 ± 0.001	0.906 ± 0.001
Avg. Prob.	0.903 ± 0.006	0.926 ± 0.003	0.872 ± 0.004	0.861 ± 0.002	0.892 ± 0.002	0.884 ± 0.005	0.866 ± 0.002	0.886 ± 0.002
Max Prob.	0.901 ± 0.011	0.925 ± 0.004	0.889 ± 0.011	0.811 ± 0.016	0.889 ± 0.006	0.869 ± 0.014	0.821 ± 0.016	0.872 ± 0.009
Majority Vote	0.860 ± 0.003	0.909 ± 0.002	0.858 ± 0.004	0.846 ± 0.002	0.863 ± 0.006	0.873 ± 0.007	0.854 ± 0.004	0.866 ± 0.002
One Monitor Pooled Data	0.917 ± 0.001	0.928 ± 0.001	0.922 ± 0.001	0.863 ± 0.001	0.899 ± 0.001	0.908 ± 0.000	0.870 ± 0.001	0.901 ± 0.001
MLP								
Configuration-Matched	0.935 ± 0.002	0.936 ± 0.003	0.933 ± 0.002	0.869 ± 0.002	0.912 ± 0.004	0.912 ± 0.001	0.877 ± 0.003	0.911 ± 0.001
Avg. Prob.	0.907 ± 0.007	0.925 ± 0.004	0.889 ± 0.009	0.864 ± 0.004	0.895 ± 0.006	0.897 ± 0.005	0.871 ± 0.004	0.892 ± 0.005
Max Prob.	0.905 ± 0.006	0.921 ± 0.006	0.860 ± 0.037	0.830 ± 0.017	0.893 ± 0.007	0.865 ± 0.017	0.836 ± 0.017	0.873 ± 0.011
Majority Vote	0.879 ± 0.009	0.906 ± 0.005	0.878 ± 0.009	0.847 ± 0.005	0.874 ± 0.009	0.886 ± 0.005	0.854 ± 0.005	0.875 ± 0.005
One Monitor Pooled Data	0.927 ± 0.001	0.932 ± 0.001	0.923 ± 0.001	0.876 ± 0.001	0.913 ± 0.001	0.918 ± 0.002	0.884 ± 0.002	0.910 ± 0.000

Table 19: **Qwen2.5-Omni-7B - $\mathcal{O}(n)$ and $\mathcal{O}(2^n)$:** Comparison of AUC using linear and MLP probes. The $\mathcal{O}(n)$ section uses Text, Image, and Audio training configurations; the $\mathcal{O}(2^n)$ section uses all seven modality combinations. Entries report mean ± standard deviation across five seeds, rounded to three decimals. The final column averages all seven test modalities.

Ensemble Method ↓ / Test →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
$\mathcal{O}(n)$: Text, Image, and Audio training								
Linear								
Avg. Prob.	0.930 ± 0.003	0.944 ± 0.001	0.897 ± 0.005	0.824 ± 0.004	0.886 ± 0.002	0.855 ± 0.004	0.834 ± 0.003	0.881 ± 0.002
Compositional Avg. Prob.	0.933 ± 0.003	0.944 ± 0.001	0.911 ± 0.005	0.821 ± 0.007	0.879 ± 0.003	0.873 ± 0.004	0.834 ± 0.003	0.885 ± 0.002
Max Prob.	0.932 ± 0.003	0.943 ± 0.002	0.911 ± 0.005	0.820 ± 0.007	0.892 ± 0.003	0.819 ± 0.016	0.829 ± 0.006	0.878 ± 0.005
Majority Vote	0.879 ± 0.006	0.908 ± 0.003	0.869 ± 0.007	0.774 ± 0.008	0.842 ± 0.007	0.834 ± 0.003	0.796 ± 0.006	0.843 ± 0.003
One Monitor Pooled Data	0.936 ± 0.001	0.943 ± 0.001	0.924 ± 0.002	0.809 ± 0.004	0.890 ± 0.003	0.890 ± 0.003	0.819 ± 0.003	0.887 ± 0.002
MLP								
Avg. Prob.	0.931 ± 0.003	0.945 ± 0.001	0.908 ± 0.003	0.823 ± 0.005	0.899 ± 0.002	0.863 ± 0.012	0.832 ± 0.005	0.886 ± 0.003
Compositional Avg. Prob.	0.937 ± 0.003	0.946 ± 0.002	0.918 ± 0.003	0.813 ± 0.009	0.894 ± 0.003	0.882 ± 0.006	0.832 ± 0.005	0.889 ± 0.002
Max Prob.	0.937 ± 0.003	0.944 ± 0.002	0.906 ± 0.006	0.816 ± 0.005	0.899 ± 0.002	0.842 ± 0.020	0.825 ± 0.004	0.881 ± 0.005
Majority Vote	0.884 ± 0.008	0.902 ± 0.005	0.882 ± 0.004	0.740 ± 0.008	0.871 ± 0.003	0.830 ± 0.011	0.762 ± 0.010	0.839 ± 0.004
One Monitor Pooled Data	0.941 ± 0.002	0.945 ± 0.001	0.927 ± 0.002	0.819 ± 0.005	0.902 ± 0.003	0.897 ± 0.005	0.830 ± 0.004	0.894 ± 0.002
$\mathcal{O}(2^n)$: All seven training configurations								
Linear								
Configuration-Matched	0.933 ± 0.003	0.944 ± 0.001	0.911 ± 0.005	0.882 ± 0.002	0.917 ± 0.001	0.919 ± 0.000	0.884 ± 0.002	0.913 ± 0.001
Avg. Prob.	0.923 ± 0.002	0.942 ± 0.001	0.905 ± 0.003	0.874 ± 0.002	0.904 ± 0.001	0.897 ± 0.002	0.880 ± 0.001	0.904 ± 0.001
Max Prob.	0.930 ± 0.004	0.941 ± 0.001	0.912 ± 0.006	0.860 ± 0.005	0.899 ± 0.003	0.864 ± 0.017	0.863 ± 0.006	0.896 ± 0.005
Majority Vote	0.882 ± 0.005	0.916 ± 0.002	0.894 ± 0.003	0.850 ± 0.002	0.891 ± 0.002	0.886 ± 0.002	0.862 ± 0.005	0.883 ± 0.001
One Monitor Pooled Data	0.933 ± 0.002	0.940 ± 0.001	0.923 ± 0.002	0.869 ± 0.002	0.917 ± 0.001	0.921 ± 0.001	0.874 ± 0.002	0.911 ± 0.002
MLP								
Configuration-Matched	0.937 ± 0.003	0.946 ± 0.002	0.918 ± 0.003	0.881 ± 0.004	0.918 ± 0.003	0.923 ± 0.002	0.887 ± 0.002	0.916 ± 0.001
Avg. Prob.	0.921 ± 0.004	0.944 ± 0.002	0.908 ± 0.002	0.875 ± 0.002	0.910 ± 0.002	0.904 ± 0.005	0.882 ± 0.002	0.906 ± 0.002
Max Prob.	0.930 ± 0.006	0.943 ± 0.002	0.910 ± 0.005	0.870 ± 0.003	0.905 ± 0.002	0.897 ± 0.011	0.879 ± 0.003	0.905 ± 0.003
Majority Vote	0.885 ± 0.006	0.915 ± 0.005	0.895 ± 0.002	0.848 ± 0.005	0.896 ± 0.003	0.894 ± 0.007	0.864 ± 0.005	0.885 ± 0.002
One Monitor Pooled Data	0.934 ± 0.003	0.940 ± 0.003	0.927 ± 0.004	0.884 ± 0.004	0.922 ± 0.003	0.925 ± 0.003	0.889 ± 0.004	0.917 ± 0.003

Table 20: **Qwen3-Omni-30B - $\mathcal{O}(n)$ and $\mathcal{O}(2^n)$:** Comparison of AUC using linear and MLP probes. The $\mathcal{O}(n)$ section uses Text, Image, and Audio training configurations; the $\mathcal{O}(2^n)$ section uses all seven modality combinations. Entries report mean ± standard deviation across five seeds, rounded to three decimals. The final column averages all seven test modalities.

Ensemble Method ↓ / Test →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
$\mathcal{O}(n)$: Text, Image, and Audio training								
Linear								
Avg. Prob.	0.926 ± 0.004	0.885 ± 0.006	0.875 ± 0.018	0.854 ± 0.014	0.915 ± 0.002	0.767 ± 0.020	0.845 ± 0.008	0.867 ± 0.004
Compositional Avg. Prob.	0.942 ± 0.001	0.929 ± 0.002	0.934 ± 0.001	0.853 ± 0.011	0.914 ± 0.003	0.814 ± 0.016	0.845 ± 0.008	0.890 ± 0.003
Max Prob.	0.922 ± 0.005	0.833 ± 0.011	0.908 ± 0.015	0.847 ± 0.016	0.904 ± 0.006	0.748 ± 0.040	0.836 ± 0.008	0.857 ± 0.002
Majority Vote	0.879 ± 0.005	0.845 ± 0.005	0.839 ± 0.020	0.705 ± 0.046	0.882 ± 0.004	0.651 ± 0.025	0.713 ± 0.037	0.788 ± 0.015
One Monitor Pooled Data	0.936 ± 0.001	0.928 ± 0.002	0.935 ± 0.001	0.857 ± 0.002	0.916 ± 0.001	0.875 ± 0.004	0.850 ± 0.001	0.899 ± 0.001
MLP								
Avg. Prob.	0.929 ± 0.005	0.898 ± 0.008	0.896 ± 0.016	0.851 ± 0.008	0.920 ± 0.003	0.748 ± 0.039	0.842 ± 0.007	0.869 ± 0.007
Compositional Avg. Prob.	0.943 ± 0.003	0.932 ± 0.003	0.938 ± 0.002	0.852 ± 0.008	0.919 ± 0.003	0.769 ± 0.018	0.842 ± 0.007	0.885 ± 0.001
Max Prob.	0.927 ± 0.007	0.896 ± 0.014	0.859 ± 0.027	0.841 ± 0.010	0.906 ± 0.011	0.729 ± 0.055	0.830 ± 0.008	0.856 ± 0.009
Majority Vote	0.897 ± 0.008	0.856 ± 0.008	0.866 ± 0.022	0.715 ± 0.045	0.893 ± 0.004	0.689 ± 0.022	0.737 ± 0.037	0.808 ± 0.010
One Monitor Pooled Data	0.941 ± 0.002	0.930 ± 0.006	0.938 ± 0.002	0.854 ± 0.010	0.920 ± 0.002	0.829 ± 0.008	0.846 ± 0.007	0.894 ± 0.003
$\mathcal{O}(2^n)$: All seven training configurations								
Linear								
Configuration-Matched	0.942 ± 0.001	0.929 ± 0.002	0.934 ± 0.001	0.902 ± 0.001	0.926 ± 0.001	0.908 ± 0.001	0.889 ± 0.002	0.918 ± 0.001
Avg. Prob.	0.926 ± 0.003	0.906 ± 0.003	0.885 ± 0.007	0.897 ± 0.002	0.920 ± 0.001	0.869 ± 0.009	0.889 ± 0.001	0.899 ± 0.002
Max Prob.	0.914 ± 0.004	0.855 ± 0.010	0.907 ± 0.014	0.885 ± 0.011	0.895 ± 0.009	0.829 ± 0.044	0.881 ± 0.007	0.881 ± 0.006
Majority Vote	0.903 ± 0.007	0.880 ± 0.006	0.873 ± 0.011	0.871 ± 0.004	0.908 ± 0.003	0.849 ± 0.010	0.871 ± 0.002	0.879 ± 0.003
One Monitor Pooled Data	0.935 ± 0.001	0.929 ± 0.001	0.933 ± 0.001	0.892 ± 0.001	0.923 ± 0.001	0.910 ± 0.001	0.886 ± 0.001	0.916 ± 0.001
MLP								
Configuration-Matched	0.943 ± 0.003	0.932 ± 0.003	0.938 ± 0.002	0.908 ± 0.002	0.929 ± 0.002	0.913 ± 0.002	0.895 ± 0.002	0.923 ± 0.001
Avg. Prob.	0.930 ± 0.004	0.913 ± 0.006	0.890 ± 0.015	0.898 ± 0.003	0.921 ± 0.003	0.877 ± 0.011	0.892 ± 0.002	0.903 ± 0.004
Max Prob.	0.927 ± 0.008	0.905 ± 0.009	0.838 ± 0.015	0.873 ± 0.030	0.906 ± 0.011	0.841 ± 0.034	0.870 ± 0.024	0.880 ± 0.011
Majority Vote	0.910 ± 0.009	0.893 ± 0.007	0.877 ± 0.018	0.871 ± 0.004	0.908 ± 0.004	0.853 ± 0.014	0.872 ± 0.005	0.883 ± 0.004
One Monitor Pooled Data	0.934 ± 0.002	0.932 ± 0.002	0.932 ± 0.003	0.904 ± 0.001	0.925 ± 0.002	0.917 ± 0.001	0.898 ± 0.001	0.920 ± 0.001

Table 21: **Phi4-mm - $\mathcal{O}(n)$ and $\mathcal{O}(2^n)$:** Comparison of AUC using linear and MLP probes. The $\mathcal{O}(n)$ section uses Text, Image, and Audio training configurations; the $\mathcal{O}(2^n)$ section uses all seven modality combinations. Entries report mean ± standard deviation across five seeds, rounded to three decimals. The final column averages all seven test modalities.

Ensemble Method ↓ / Test →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
$\mathcal{O}(n)$: Text, Image, and Audio training								
Linear								
Avg. Prob.	0.935 ± 0.001	0.917 ± 0.004	0.907 ± 0.003	0.766 ± 0.004	0.868 ± 0.004	0.735 ± 0.010	0.766 ± 0.004	0.842 ± 0.003
Compositional Avg. Prob.	0.940 ± 0.002	0.922 ± 0.002	0.935 ± 0.001	0.764 ± 0.004	0.857 ± 0.003	0.741 ± 0.010	0.766 ± 0.004	0.846 ± 0.002
Max Prob.	0.936 ± 0.002	0.917 ± 0.004	0.929 ± 0.002	0.763 ± 0.005	0.857 ± 0.003	0.734 ± 0.010	0.763 ± 0.005	0.843 ± 0.002
Majority Vote	0.885 ± 0.006	0.858 ± 0.005	0.859 ± 0.003	0.674 ± 0.010	0.805 ± 0.021	0.639 ± 0.008	0.669 ± 0.010	0.770 ± 0.003
One Monitor Pooled Data	0.941 ± 0.001	0.922 ± 0.003	0.934 ± 0.001	0.756 ± 0.003	0.856 ± 0.004	0.745 ± 0.008	0.761 ± 0.003	0.845 ± 0.003
MLP								
Avg. Prob.	0.936 ± 0.003	0.918 ± 0.004	0.915 ± 0.007	0.743 ± 0.009	0.874 ± 0.004	0.745 ± 0.022	0.747 ± 0.009	0.840 ± 0.004
Compositional Avg. Prob.	0.944 ± 0.002	0.924 ± 0.003	0.935 ± 0.003	0.735 ± 0.010	0.866 ± 0.002	0.755 ± 0.018	0.747 ± 0.009	0.844 ± 0.003
Max Prob.	0.933 ± 0.007	0.918 ± 0.002	0.930 ± 0.004	0.736 ± 0.008	0.860 ± 0.005	0.740 ± 0.024	0.740 ± 0.008	0.837 ± 0.003
Majority Vote	0.897 ± 0.010	0.862 ± 0.008	0.868 ± 0.002	0.622 ± 0.033	0.806 ± 0.018	0.632 ± 0.037	0.629 ± 0.033	0.759 ± 0.017
One Monitor Pooled Data	0.943 ± 0.002	0.923 ± 0.007	0.935 ± 0.002	0.741 ± 0.009	0.862 ± 0.007	0.753 ± 0.015	0.746 ± 0.009	0.843 ± 0.005
$\mathcal{O}(2^n)$: All seven training configurations								
Linear								
Configuration-Matched	0.940 ± 0.002	0.922 ± 0.002	0.935 ± 0.001	0.854 ± 0.001	0.899 ± 0.003	0.899 ± 0.002	0.858 ± 0.001	0.901 ± 0.000
Avg. Prob.	0.936 ± 0.001	0.915 ± 0.003	0.918 ± 0.002	0.847 ± 0.002	0.884 ± 0.001	0.860 ± 0.003	0.851 ± 0.002	0.887 ± 0.001
Max Prob.	0.935 ± 0.002	0.911 ± 0.003	0.924 ± 0.001	0.843 ± 0.003	0.873 ± 0.003	0.863 ± 0.005	0.848 ± 0.002	0.885 ± 0.002
Majority Vote	0.898 ± 0.010	0.878 ± 0.006	0.898 ± 0.003	0.819 ± 0.003	0.852 ± 0.005	0.845 ± 0.003	0.831 ± 0.001	0.860 ± 0.002
One Monitor Pooled Data	0.935 ± 0.001	0.916 ± 0.002	0.931 ± 0.001	0.842 ± 0.002	0.888 ± 0.001	0.891 ± 0.002	0.846 ± 0.002	0.892 ± 0.001
MLP								
Configuration-Matched	0.944 ± 0.002	0.924 ± 0.003	0.935 ± 0.003	0.850 ± 0.003	0.896 ± 0.003	0.898 ± 0.002	0.856 ± 0.003	0.900 ± 0.000
Avg. Prob.	0.924 ± 0.004	0.898 ± 0.009	0.903 ± 0.006	0.841 ± 0.002	0.882 ± 0.003	0.862 ± 0.008	0.846 ± 0.002	0.879 ± 0.004
Max Prob.	0.915 ± 0.021	0.890 ± 0.016	0.913 ± 0.016	0.835 ± 0.006	0.860 ± 0.027	0.866 ± 0.011	0.844 ± 0.004	0.875 ± 0.012
Majority Vote	0.906 ± 0.006	0.873 ± 0.013	0.889 ± 0.007	0.810 ± 0.006	0.865 ± 0.003	0.842 ± 0.006	0.821 ± 0.004	0.858 ± 0.004
One Monitor Pooled Data	0.938 ± 0.003	0.924 ± 0.004	0.929 ± 0.003	0.859 ± 0.003	0.903 ± 0.004	0.901 ± 0.003	0.863 ± 0.003	0.902 ± 0.002

Table 22: **Interactive-Omni-8B - $\mathcal{O}(n)$ and $\mathcal{O}(2^n)$:** Comparison of AUC using linear and MLP probes. The $\mathcal{O}(n)$ section uses Text, Image, and Audio training configurations; the $\mathcal{O}(2^n)$ section uses all seven modality combinations. Entries report mean $\pm$ standard deviation across five seeds, rounded to three decimals. The final column averages all seven test modalities.

Ensemble Method ↓ / Test →	Text	Image	Audio	Text+Image	Text+Audio	Audio+Image	Text+Audio+Image	Avg.
$\mathcal{O}(n)$: Text, Image, and Audio training								
Linear								
Avg. Prob.	0.926 ± 0.003	0.919 ± 0.002	0.920 ± 0.003	0.833 ± 0.003	0.899 ± 0.001	0.903 ± 0.003	0.867 ± 0.002	0.895 ± 0.001
Compositional Avg. Prob.	0.941 ± 0.001	0.930 ± 0.001	0.929 ± 0.003	0.847 ± 0.004	0.890 ± 0.002	0.903 ± 0.003	0.867 ± 0.002	0.901 ± 0.001
Max Prob.	0.932 ± 0.003	0.911 ± 0.006	0.931 ± 0.002	0.819 ± 0.005	0.888 ± 0.001	0.906 ± 0.004	0.856 ± 0.004	0.892 ± 0.003
Majority Vote	0.874 ± 0.003	0.891 ± 0.003	0.868 ± 0.003	0.746 ± 0.009	0.865 ± 0.003	0.861 ± 0.007	0.827 ± 0.005	0.847 ± 0.002
One Monitor Pooled Data	0.937 ± 0.002	0.930 ± 0.001	0.932 ± 0.001	0.851 ± 0.004	0.902 ± 0.002	0.911 ± 0.001	0.875 ± 0.002	0.905 ± 0.002
MLP								
Avg. Prob.	0.921 ± 0.004	0.927 ± 0.003	0.923 ± 0.004	0.849 ± 0.010	0.901 ± 0.004	0.903 ± 0.003	0.872 ± 0.004	0.899 ± 0.001
Compositional Avg. Prob.	0.941 ± 0.002	0.926 ± 0.007	0.932 ± 0.002	0.842 ± 0.011	0.893 ± 0.006	0.904 ± 0.003	0.872 ± 0.004	0.901 ± 0.003
Max Prob.	0.924 ± 0.007	0.922 ± 0.003	0.931 ± 0.003	0.837 ± 0.009	0.887 ± 0.007	0.902 ± 0.004	0.861 ± 0.005	0.895 ± 0.003
Majority Vote	0.878 ± 0.006	0.895 ± 0.006	0.873 ± 0.009	0.771 ± 0.019	0.867 ± 0.012	0.858 ± 0.010	0.834 ± 0.007	0.854 ± 0.003
One Monitor Pooled Data	0.939 ± 0.004	0.931 ± 0.002	0.930 ± 0.004	0.849 ± 0.007	0.903 ± 0.004	0.910 ± 0.005	0.876 ± 0.002	0.905 ± 0.002
$\mathcal{O}(2^n)$: All seven training configurations								
Linear								
Configuration-Matched	0.941 ± 0.001	0.930 ± 0.001	0.929 ± 0.003	0.904 ± 0.002	0.915 ± 0.002	0.916 ± 0.001	0.905 ± 0.001	0.920 ± 0.001
Avg. Prob.	0.925 ± 0.002	0.917 ± 0.003	0.917 ± 0.003	0.890 ± 0.002	0.911 ± 0.001	0.914 ± 0.001	0.899 ± 0.002	0.910 ± 0.001
Max Prob.	0.926 ± 0.002	0.910 ± 0.003	0.922 ± 0.002	0.889 ± 0.004	0.904 ± 0.002	0.911 ± 0.002	0.889 ± 0.002	0.907 ± 0.001
Majority Vote	0.881 ± 0.003	0.900 ± 0.004	0.898 ± 0.003	0.858 ± 0.004	0.899 ± 0.002	0.900 ± 0.003	0.880 ± 0.002	0.888 ± 0.001
One Monitor Pooled Data	0.934 ± 0.001	0.924 ± 0.003	0.925 ± 0.002	0.891 ± 0.003	0.910 ± 0.003	0.913 ± 0.002	0.903 ± 0.002	0.914 ± 0.002
MLP								
Configuration-Matched	0.941 ± 0.002	0.926 ± 0.007	0.932 ± 0.002	0.900 ± 0.006	0.915 ± 0.004	0.920 ± 0.003	0.907 ± 0.005	0.920 ± 0.003
Avg. Prob.	0.926 ± 0.004	0.923 ± 0.004	0.919 ± 0.002	0.893 ± 0.003	0.915 ± 0.002	0.917 ± 0.002	0.902 ± 0.003	0.913 ± 0.001
Max Prob.	0.925 ± 0.006	0.919 ± 0.005	0.916 ± 0.008	0.888 ± 0.003	0.904 ± 0.003	0.902 ± 0.010	0.889 ± 0.004	0.906 ± 0.003
Majority Vote	0.891 ± 0.009	0.902 ± 0.003	0.902 ± 0.004	0.870 ± 0.005	0.898 ± 0.006	0.905 ± 0.003	0.884 ± 0.005	0.893 ± 0.001
One Monitor Pooled Data	0.937 ± 0.003	0.924 ± 0.003	0.923 ± 0.004	0.898 ± 0.003	0.917 ± 0.002	0.918 ± 0.003	0.907 ± 0.003	0.918 ± 0.002

C.2 FULL RESULTS FOR ABLATION STUDIES

Table 23 reports the full results for the ablations presented in Section 4.3. The full $\mathcal{O}(1)$ method achieves the highest average AUC across all evaluated models and probe architectures.

Table 23: **Ablations of safety, semantic, and modality covariance.** AUC for our implementation of the new $\mathcal{O}(1)$ method and four covariance ablations across five omni-models on WGM-Omni. Entries report mean ± sample standard deviation across five seeds, rounded to three decimals; Avg. averages all seven evaluation configurations.

Monitor	Method ↓ / Eval →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	w/o $\mathbf{S}_{\text{safety}}$	0.896 ± 0.000	0.906 ± 0.001	0.776 ± 0.005	0.786 ± 0.001	0.856 ± 0.000	0.803 ± 0.002	0.788 ± 0.001	0.830 ± 0.001
	w/o $\mathbf{S}_{\text{modality}}$	0.922 ± 0.001	0.916 ± 0.002	0.611 ± 0.027	0.791 ± 0.009	0.860 ± 0.004	0.701 ± 0.016	0.779 ± 0.009	0.797 ± 0.009
	w/o $\mathbf{S}_{\text{semantic}}$	0.886 ± 0.011	0.881 ± 0.013	0.752 ± 0.050	0.747 ± 0.004	0.836 ± 0.006	0.754 ± 0.039	0.750 ± 0.004	0.801 ± 0.008
	General $\mathbf{S}_{\text{semantic}}$	0.881 ± 0.001	0.890 ± 0.003	**0.807 ± 0.007**	0.763 ± 0.002	0.839 ± 0.002	0.773 ± 0.004	0.763 ± 0.002	0.817 ± 0.002
	Ours	**0.923 ± 0.006**	**0.932 ± 0.001**	0.802 ± 0.006	**0.801 ± 0.004**	**0.887 ± 0.003**	**0.816 ± 0.005**	**0.803 ± 0.005**	**0.852 ± 0.002**
MLP	w/o $\mathbf{S}_{\text{safety}}$	0.925 ± 0.002	0.926 ± 0.002	0.838 ± 0.005	0.795 ± 0.006	0.883 ± 0.001	0.823 ± 0.006	0.798 ± 0.006	0.855 ± 0.003
	w/o $\mathbf{S}_{\text{modality}}$	**0.933 ± 0.003**	0.914 ± 0.006	0.687 ± 0.037	0.772 ± 0.011	0.874 ± 0.005	0.704 ± 0.020	0.762 ± 0.009	0.806 ± 0.008
	w/o $\mathbf{S}_{\text{semantic}}$	0.896 ± 0.012	0.878 ± 0.012	0.757 ± 0.035	0.729 ± 0.032	0.835 ± 0.010	0.759 ± 0.033	0.734 ± 0.031	0.798 ± 0.016
	General $\mathbf{S}_{\text{semantic}}$	0.911 ± 0.003	0.911 ± 0.007	0.832 ± 0.010	0.785 ± 0.003	0.860 ± 0.004	0.834 ± 0.004	0.789 ± 0.003	0.846 ± 0.004
	Ours	**0.933 ± 0.003**	**0.934 ± 0.001**	**0.844 ± 0.012**	**0.811 ± 0.007**	**0.901 ± 0.002**	**0.844 ± 0.007**	**0.814 ± 0.008**	**0.869 ± 0.001**
Qwen2.5-Omni-7B									
Linear	w/o $\mathbf{S}_{\text{safety}}$	0.898 ± 0.001	0.908 ± 0.001	0.851 ± 0.002	0.778 ± 0.001	0.851 ± 0.001	0.798 ± 0.002	0.780 ± 0.001	0.838 ± 0.001
	w/o $\mathbf{S}_{\text{modality}}$	**0.927 ± 0.001**	0.929 ± 0.002	**0.905 ± 0.002**	0.793 ± 0.004	0.877 ± 0.002	0.738 ± 0.005	0.795 ± 0.004	0.852 ± 0.002
	w/o $\mathbf{S}_{\text{semantic}}$	0.899 ± 0.012	0.902 ± 0.011	0.821 ± 0.025	0.753 ± 0.012	0.841 ± 0.018	0.753 ± 0.029	0.756 ± 0.015	0.818 ± 0.017
	General $\mathbf{S}_{\text{semantic}}$	0.902 ± 0.001	0.919 ± 0.001	0.867 ± 0.001	0.776 ± 0.002	0.865 ± 0.001	0.806 ± 0.002	0.778 ± 0.002	0.845 ± 0.001
	Ours	0.924 ± 0.001	**0.936 ± 0.002**	0.889 ± 0.007	**0.824 ± 0.002**	**0.888 ± 0.003**	**0.831 ± 0.009**	**0.824 ± 0.001**	**0.874 ± 0.003**
MLP	w/o $\mathbf{S}_{\text{safety}}$	0.926 ± 0.002	0.924 ± 0.001	0.900 ± 0.002	0.799 ± 0.004	0.889 ± 0.003	**0.853 ± 0.007**	0.803 ± 0.005	0.870 ± 0.003
	w/o $\mathbf{S}_{\text{modality}}$	**0.934 ± 0.001**	**0.938 ± 0.002**	**0.909 ± 0.005**	0.791 ± 0.011	0.882 ± 0.002	0.735 ± 0.021	0.791 ± 0.012	0.854 ± 0.005
	w/o $\mathbf{S}_{\text{semantic}}$	0.903 ± 0.010	0.902 ± 0.009	0.828 ± 0.017	0.749 ± 0.013	0.845 ± 0.014	0.754 ± 0.024	0.751 ± 0.014	0.819 ± 0.010
	General $\mathbf{S}_{\text{semantic}}$	0.921 ± 0.001	0.929 ± 0.001	0.884 ± 0.002	0.778 ± 0.004	0.885 ± 0.001	0.829 ± 0.003	0.780 ± 0.004	0.858 ± 0.002
	Ours	0.932 ± 0.004	**0.938 ± 0.002**	0.904 ± 0.005	**0.816 ± 0.005**	**0.899 ± 0.006**	0.846 ± 0.012	**0.819 ± 0.005**	**0.879 ± 0.005**
Qwen3-Omni-30B									
Linear	w/o $\mathbf{S}_{\text{safety}}$	0.898 ± 0.000	0.864 ± 0.001	0.738 ± 0.004	0.826 ± 0.002	0.878 ± 0.001	0.757 ± 0.003	0.814 ± 0.002	0.825 ± 0.001
	w/o $\mathbf{S}_{\text{modality}}$	**0.937 ± 0.001**	0.856 ± 0.017	0.769 ± 0.033	0.804 ± 0.007	0.900 ± 0.001	0.589 ± 0.018	0.785 ± 0.007	0.806 ± 0.009
	w/o $\mathbf{S}_{\text{semantic}}$	0.923 ± 0.004	0.867 ± 0.012	0.745 ± 0.014	0.803 ± 0.006	0.889 ± 0.005	0.658 ± 0.016	0.781 ± 0.006	0.809 ± 0.004
	General $\mathbf{S}_{\text{semantic}}$	0.911 ± 0.001	0.852 ± 0.002	**0.850 ± 0.002**	0.801 ± 0.002	0.885 ± 0.001	0.770 ± 0.003	0.788 ± 0.002	0.837 ± 0.002
	Ours	0.935 ± 0.001	**0.918 ± 0.000**	0.833 ± 0.014	**0.864 ± 0.001**	**0.911 ± 0.002**	**0.782 ± 0.008**	**0.847 ± 0.001**	**0.870 ± 0.002**
MLP	w/o $\mathbf{S}_{\text{safety}}$	0.927 ± 0.002	0.871 ± 0.013	0.843 ± 0.007	0.827 ± 0.007	0.907 ± 0.002	0.778 ± 0.010	0.817 ± 0.009	0.853 ± 0.006
	w/o $\mathbf{S}_{\text{modality}}$	**0.942 ± 0.001**	0.884 ± 0.025	0.764 ± 0.027	0.805 ± 0.018	0.904 ± 0.003	0.598 ± 0.030	0.791 ± 0.019	0.812 ± 0.012
	w/o $\mathbf{S}_{\text{semantic}}$	0.926 ± 0.004	0.768 ± 0.059	0.742 ± 0.026	0.693 ± 0.030	0.885 ± 0.001	0.589 ± 0.011	0.679 ± 0.033	0.755 ± 0.014
	General $\mathbf{S}_{\text{semantic}}$	0.932 ± 0.001	0.837 ± 0.006	**0.872 ± 0.004**	0.797 ± 0.013	0.899 ± 0.003	0.734 ± 0.022	0.782 ± 0.013	0.836 ± 0.008
	Ours	0.937 ± 0.002	**0.912 ± 0.005**	0.860 ± 0.005	**0.846 ± 0.006**	**0.917 ± 0.001**	**0.785 ± 0.014**	**0.828 ± 0.005**	**0.869 ± 0.002**
Phi4-mm 5.6B									
Linear	w/o $\mathbf{S}_{\text{safety}}$	0.868 ± 0.001	0.808 ± 0.002	0.745 ± 0.004	0.666 ± 0.002	0.758 ± 0.002	0.615 ± 0.002	0.671 ± 0.001	0.733 ± 0.001
	w/o $\mathbf{S}_{\text{modality}}$	**0.935 ± 0.001**	0.851 ± 0.004	0.607 ± 0.033	0.705 ± 0.007	0.820 ± 0.003	0.619 ± 0.009	0.698 ± 0.007	0.748 ± 0.006
	w/o $\mathbf{S}_{\text{semantic}}$	0.918 ± 0.003	**0.865 ± 0.001**	0.769 ± 0.020	0.721 ± 0.010	0.813 ± 0.010	0.686 ± 0.013	0.721 ± 0.009	0.785 ± 0.006
	General $\mathbf{S}_{\text{semantic}}$	0.924 ± 0.001	0.855 ± 0.001	**0.872 ± 0.002**	0.701 ± 0.002	0.823 ± 0.002	**0.690 ± 0.002**	0.703 ± 0.002	0.795 ± 0.001
	Ours	0.931 ± 0.004	0.863 ± 0.011	0.845 ± 0.006	**0.725 ± 0.012**	**0.827 ± 0.007**	0.671 ± 0.012	**0.724 ± 0.012**	**0.798 ± 0.009**
MLP	w/o $\mathbf{S}_{\text{safety}}$	0.928 ± 0.001	0.856 ± 0.003	0.876 ± 0.005	0.688 ± 0.009	0.842 ± 0.005	0.662 ± 0.009	0.693 ± 0.009	0.792 ± 0.005
	w/o $\mathbf{S}_{\text{modality}}$	0.939 ± 0.002	0.827 ± 0.029	0.757 ± 0.055	0.646 ± 0.034	0.813 ± 0.012	0.585 ± 0.028	0.638 ± 0.034	0.744 ± 0.013
	w/o $\mathbf{S}_{\text{semantic}}$	0.920 ± 0.004	0.845 ± 0.011	0.762 ± 0.049	0.697 ± 0.015	0.803 ± 0.013	0.673 ± 0.014	0.695 ± 0.017	0.771 ± 0.009
	General $\mathbf{S}_{\text{semantic}}$	0.929 ± 0.001	0.849 ± 0.007	**0.881 ± 0.005**	0.703 ± 0.008	0.827 ± 0.002	**0.713 ± 0.003**	0.709 ± 0.008	0.802 ± 0.003
	Ours	**0.943 ± 0.001**	**0.895 ± 0.006**	0.860 ± 0.005	**0.742 ± 0.008**	**0.877 ± 0.004**	0.706 ± 0.015	**0.740 ± 0.009**	**0.823 ± 0.006**
Interactive-Omni-8B									
Linear	w/o $\mathbf{S}_{\text{safety}}$	0.909 ± 0.001	0.905 ± 0.001	0.798 ± 0.003	0.787 ± 0.001	0.842 ± 0.002	0.789 ± 0.003	0.809 ± 0.002	0.834 ± 0.001
	w/o $\mathbf{S}_{\text{modality}}$	**0.934 ± 0.001**	0.906 ± 0.004	0.852 ± 0.004	0.833 ± 0.002	0.878 ± 0.002	0.821 ± 0.006	0.834 ± 0.002	0.865 ± 0.001
	w/o $\mathbf{S}_{\text{semantic}}$	0.905 ± 0.001	0.886 ± 0.004	0.738 ± 0.026	0.786 ± 0.016	0.835 ± 0.009	0.761 ± 0.021	0.796 ± 0.013	0.815 ± 0.011
	General $\mathbf{S}_{\text{semantic}}$	0.913 ± 0.001	0.894 ± 0.001	0.843 ± 0.002	0.821 ± 0.001	0.866 ± 0.001	0.820 ± 0.002	0.831 ± 0.001	0.855 ± 0.001
	Ours	0.929 ± 0.005	**0.926 ± 0.002**	**0.858 ± 0.012**	**0.852 ± 0.003**	**0.882 ± 0.008**	**0.851 ± 0.006**	**0.856 ± 0.004**	**0.879 ± 0.006**
MLP	w/o $\mathbf{S}_{\text{safety}}$	0.934 ± 0.003	0.922 ± 0.002	**0.875 ± 0.005**	0.827 ± 0.003	0.881 ± 0.003	0.849 ± 0.004	0.853 ± 0.002	0.877 ± 0.001
	w/o $\mathbf{S}_{\text{modality}}$	**0.940 ± 0.002**	0.904 ± 0.008	0.843 ± 0.018	0.828 ± 0.009	0.879 ± 0.003	0.812 ± 0.027	0.837 ± 0.007	0.863 ± 0.009
	w/o $\mathbf{S}_{\text{semantic}}$	0.916 ± 0.002	0.882 ± 0.006	0.765 ± 0.054	0.785 ± 0.017	0.842 ± 0.014	0.780 ± 0.016	0.800 ± 0.014	0.824 ± 0.014
	General $\mathbf{S}_{\text{semantic}}$	0.932 ± 0.001	0.904 ± 0.002	0.856 ± 0.005	0.837 ± 0.002	0.885 ± 0.002	0.852 ± 0.002	0.852 ± 0.002	0.874 ± 0.002
	Ours	0.937 ± 0.002	**0.929 ± 0.003**	0.874 ± 0.005	**0.853 ± 0.009**	**0.895 ± 0.005**	**0.860 ± 0.005**	**0.865 ± 0.007**	**0.887 ± 0.005**

C.3 Additional analysis for three supervision regimes' comparison

We report the full AUC, accuracy, and F1 results of the three supervision regimes in Tables 29–34. Overall, accuracy and F1 largely preserve the supervision-scaling trends observed with AUC, with the larger modality gap particularly pronounced on WGM-Omni. Moreover, augmenting the $\mathcal{O}(n)$ setting with text+image+audio supervision substantially narrows the gap to specialized $\mathcal{O}(2^n)$ monitoring. The proposed $\mathcal{O}(1)$ subspace method also generally improves over the original $\mathcal{O}(1)$ baseline, although a small number of regressions occur in accuracy and F1. Since these metrics use a fixed threshold of 0.5, whereas the corresponding AUC results improve across all settings, these regressions are consistent with changes in score calibration or the optimal decision threshold rather than degraded ranking performance. We also visualize the original and projected text, image, and audio representations using PCA in Figure 10. After projection, the representations exhibit substantially less separation by modality.

C.3.1 Comparison with direct PCA

Beyond the original $\mathcal{O}(1)$ baseline, we compare our method with direct PCA using a dimension-matched subspace. As shown in Table 24, our method substantially outperforms direct PCA, indicating that the gains cannot be attributed to dimensionality reduction alone.

Table 24: **Comparison with dimension-matched PCA.** Direct PCA uses the same subspace dimension as our method.

Monitor	Method	Qwen2.5-Omni-3B	Qwen2.5-Omni-7B	Qwen3-Omni-30B	Phi4-mm	Interactive-Omni-8B
Linear	Direct PCA	0.787 $\pm$ 0.001	0.787 $\pm$ 0.003	0.799 $\pm$ 0.008	0.698 $\pm$ 0.008	0.849 $\pm$ 0.004
	Ours	**0.852** $\pm$ **0.002**	**0.874** $\pm$ **0.003**	**0.870** $\pm$ **0.002**	**0.798** $\pm$ **0.009**	**0.879** $\pm$ **0.006**
MLP	Direct PCA	0.788 $\pm$ 0.007	0.768 $\pm$ 0.009	0.810 $\pm$ 0.007	0.752 $\pm$ 0.008	0.858 $\pm$ 0.010
	Ours	**0.869** $\pm$ **0.001**	**0.879** $\pm$ **0.005**	**0.869** $\pm$ **0.002**	**0.823** $\pm$ **0.006**	**0.887** $\pm$ **0.005**

C.3.2 Cross-dataset generalization and continual updating

We show the full results of cross-dataset generalization and continual monitor updating in Table 25. Our method consistently achieves higher average AUC across all models and monitors. The gains are particularly pronounced on audio-containing configurations, where the original $\mathcal{O}(1)$ baseline exhibits larger degradation under dataset shift. These results highlight an important practical advantage of our method: general omni-modal data are required only once for subspace initialization, while subsequent adaptation to new safety distributions can rely solely on text safety data without re-estimating the subspace.

C.3.3 Case studies

Case Study: Sensitivity to Layer Selection. In the main study, we use final-layer representations as a consistent extraction setting across models. To verify that our results are not specific to this choice, we conduct a case study on `Qwen2.5-Omni-3B` and `Phi4-mm` using WGM-Omni, additionally extracting representations at the $\frac{1}{2}$ and $\frac{3}{4}$ layer positions. Together with the final layer, these settings span the mid-to-late layers of the LLM backbone, which prior work has shown to encode high-level semantic information useful for safety monitoring (Oldfield et al., 2026; Park et al., 2024). Comparing Tables 27 and 28 with the final-layer results in Table 29, we find that our main supervision-scaling findings persist across layer choices. First, the modality gap remains larger than the combination gap. Second, augmenting the $\mathcal{O}(n)$ setting with text+image+audio supervision continues to substantially narrow the gap to specialized $\mathcal{O}(2^n)$ monitoring. Third, our $\mathcal{O}(1)$ subspace method continues to improve over the original $\mathcal{O}(1)$ baseline. These results suggest that the main conclusions are not specific to final-layer representations.

Case Study: Sensitivity to the Choice of General Dataset. In the main study, we use NoRobots (Rajani et al., 2023) as the general, non-safety-related data for estimating modality variation in our $\mathcal{O}(1)$ subspace method. To assess sensitivity to this choice, we repeat the experiment using Dolly (Conover et al., 2023), another general dataset, on Aegis-Omni with `Qwen2.5-Omni-3B` and `Phi4-mm`. As shown in Table 26, both NoRobots and Dolly substantially improve over the original $\mathcal{O}(1)$ baseline across models and probe architectures, while their relative performance varies by

Table 25: **Cross-dataset generalization AUC from WGM-Omni to Aegis-Omni.** We compare the original $\mathcal{O}(1)$ baseline and our $\mathcal{O}(1)$ method before and after continual updating on Aegis text data. Entries report mean $\pm$ sample standard deviation over five seeds, rounded to three decimals.

Monitor	Method ↓ / Test →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	$\mathcal{O}(1)$	**0.787** $\pm$ **0.002**	0.829 $\pm$ 0.005	0.388 $\pm$ 0.026	**0.684** $\pm$ **0.007**	0.589 $\pm$ 0.009	0.513 $\pm$ 0.008	0.667 $\pm$ 0.007	0.637 $\pm$ 0.005
	$\mathcal{O}(1)$ **- Ours**	0.781 $\pm$ 0.018	**0.844** $\pm$ **0.003**	**0.605** $\pm$ **0.012**	0.681 $\pm$ 0.014	**0.672** $\pm$ **0.011**	**0.645** $\pm$ **0.019**	**0.674** $\pm$ **0.013**	**0.700** $\pm$ **0.003**
	$\mathcal{O}(1)$ **- Updated**	0.874 $\pm$ 0.002	0.873 $\pm$ 0.005	0.422 $\pm$ 0.028	0.728 $\pm$ 0.006	0.661 $\pm$ 0.012	0.599 $\pm$ 0.006	0.720 $\pm$ 0.006	0.696 $\pm$ 0.005
	$\mathcal{O}(1)$ **- Ours - Updated**	**0.880** $\pm$ **0.011**	**0.879** $\pm$ **0.004**	**0.673** $\pm$ **0.014**	**0.744** $\pm$ **0.013**	**0.800** $\pm$ **0.012**	**0.723** $\pm$ **0.009**	**0.740** $\pm$ **0.010**	**0.777** $\pm$ **0.002**
MLP	$\mathcal{O}(1)$	0.811 $\pm$ 0.008	0.834 $\pm$ 0.011	0.381 $\pm$ 0.029	0.678 $\pm$ 0.026	0.604 $\pm$ 0.012	0.498 $\pm$ 0.014	0.652 $\pm$ 0.018	0.637 $\pm$ 0.011
	$\mathcal{O}(1)$ **- Ours**	**0.833** $\pm$ **0.008**	**0.836** $\pm$ **0.010**	**0.589** $\pm$ **0.022**	**0.700** $\pm$ **0.015**	**0.672** $\pm$ **0.018**	**0.615** $\pm$ **0.034**	**0.690** $\pm$ **0.019**	**0.705** $\pm$ **0.011**
	$\mathcal{O}(1)$ **- Updated**	0.887 $\pm$ 0.003	**0.871** $\pm$ **0.009**	0.511 $\pm$ 0.044	0.712 $\pm$ 0.032	0.729 $\pm$ 0.025	0.629 $\pm$ 0.013	0.706 $\pm$ 0.022	0.721 $\pm$ 0.014
	$\mathcal{O}(1)$ **- Ours - Updated**	**0.895** $\pm$ **0.008**	0.863 $\pm$ 0.006	**0.652** $\pm$ **0.020**	**0.737** $\pm$ **0.020**	**0.773** $\pm$ **0.017**	**0.683** $\pm$ **0.012**	**0.730** $\pm$ **0.021**	**0.762** $\pm$ **0.008**
Qwen2.5-Omni-7B									
Linear	$\mathcal{O}(1)$	**0.816** $\pm$ **0.006**	0.856 $\pm$ 0.002	0.667 $\pm$ 0.008	**0.726** $\pm$ **0.004**	0.683 $\pm$ 0.008	0.689 $\pm$ 0.015	**0.721** $\pm$ **0.004**	0.737 $\pm$ 0.005
	$\mathcal{O}(1)$ **- Ours**	0.807 $\pm$ 0.007	**0.868** $\pm$ **0.001**	**0.700** $\pm$ **0.011**	0.720 $\pm$ 0.002	**0.721** $\pm$ **0.004**	**0.699** $\pm$ **0.009**	0.714 $\pm$ 0.002	**0.747** $\pm$ **0.004**
	$\mathcal{O}(1)$ **- Updated**	**0.888** $\pm$ **0.001**	**0.886** $\pm$ **0.002**	0.678 $\pm$ 0.010	0.749 $\pm$ 0.003	0.750 $\pm$ 0.006	**0.742** $\pm$ **0.009**	0.754 $\pm$ 0.003	0.778 $\pm$ 0.004
	$\mathcal{O}(1)$ **- Ours - Updated**	0.876 $\pm$ 0.002	0.885 $\pm$ 0.000	**0.740** $\pm$ **0.012**	**0.769** $\pm$ **0.001**	**0.800** $\pm$ **0.005**	0.741 $\pm$ 0.014	**0.763** $\pm$ **0.002**	**0.796** $\pm$ **0.005**
MLP	$\mathcal{O}(1)$	0.836 $\pm$ 0.008	0.847 $\pm$ 0.005	0.655 $\pm$ 0.031	0.696 $\pm$ 0.009	0.691 $\pm$ 0.018	0.685 $\pm$ 0.022	0.689 $\pm$ 0.009	0.728 $\pm$ 0.011
	$\mathcal{O}(1)$ **- Ours**	**0.840** $\pm$ **0.005**	**0.871** $\pm$ **0.003**	**0.715** $\pm$ **0.010**	**0.738** $\pm$ **0.011**	**0.754** $\pm$ **0.008**	**0.705** $\pm$ **0.008**	**0.734** $\pm$ **0.011**	**0.765** $\pm$ **0.006**
	$\mathcal{O}(1)$ **- Updated**	**0.902** $\pm$ **0.002**	**0.889** $\pm$ **0.001**	0.715 $\pm$ 0.020	0.729 $\pm$ 0.012	0.788 $\pm$ 0.012	**0.761** $\pm$ **0.014**	0.739 $\pm$ 0.012	0.789 $\pm$ 0.009
	$\mathcal{O}(1)$ **- Ours - Updated**	0.885 $\pm$ 0.004	0.885 $\pm$ 0.002	**0.743** $\pm$ **0.004**	**0.774** $\pm$ **0.004**	**0.809** $\pm$ **0.003**	0.733 $\pm$ 0.009	**0.770** $\pm$ **0.005**	**0.800** $\pm$ **0.002**
Qwen3-Omni-30B									
Linear	$\mathcal{O}(1)$	0.805 $\pm$ 0.006	0.710 $\pm$ 0.018	0.453 $\pm$ 0.012	0.641 $\pm$ 0.018	0.655 $\pm$ 0.005	0.456 $\pm$ 0.029	0.590 $\pm$ 0.011	0.616 $\pm$ 0.006
	$\mathcal{O}(1)$ **- Ours**	**0.843** $\pm$ **0.004**	**0.796** $\pm$ **0.015**	**0.650** $\pm$ **0.013**	**0.685** $\pm$ **0.007**	**0.751** $\pm$ **0.002**	**0.647** $\pm$ **0.009**	**0.659** $\pm$ **0.006**	**0.719** $\pm$ **0.004**
	$\mathcal{O}(1)$ **- Updated**	0.880 $\pm$ 0.005	0.749 $\pm$ 0.009	0.551 $\pm$ 0.006	0.711 $\pm$ 0.012	0.771 $\pm$ 0.006	0.499 $\pm$ 0.021	0.667 $\pm$ 0.015	0.690 $\pm$ 0.006
	$\mathcal{O}(1)$ **- Ours - Updated**	**0.884** $\pm$ **0.004**	**0.861** $\pm$ **0.004**	**0.668** $\pm$ **0.017**	**0.774** $\pm$ **0.011**	**0.805** $\pm$ **0.008**	**0.697** $\pm$ **0.017**	**0.748** $\pm$ **0.014**	**0.777** $\pm$ **0.010**
MLP	$\mathcal{O}(1)$	0.818 $\pm$ 0.015	0.737 $\pm$ 0.020	0.478 $\pm$ 0.023	0.660 $\pm$ 0.008	0.689 $\pm$ 0.008	0.464 $\pm$ 0.015	0.593 $\pm$ 0.011	0.634 $\pm$ 0.009
	$\mathcal{O}(1)$ **- Ours**	**0.858** $\pm$ **0.005**	**0.795** $\pm$ **0.017**	**0.668** $\pm$ **0.015**	**0.693** $\pm$ **0.016**	**0.762** $\pm$ **0.007**	**0.665** $\pm$ **0.010**	**0.664** $\pm$ **0.017**	**0.729** $\pm$ **0.006**
	$\mathcal{O}(1)$ **- Updated**	**0.894** $\pm$ **0.003**	0.751 $\pm$ 0.015	0.557 $\pm$ 0.011	0.704 $\pm$ 0.009	0.799 $\pm$ 0.004	0.522 $\pm$ 0.010	0.652 $\pm$ 0.013	0.697 $\pm$ 0.006
	$\mathcal{O}(1)$ **- Ours - Updated**	0.888 $\pm$ 0.001	**0.834** $\pm$ **0.015**	**0.667** $\pm$ **0.016**	**0.765** $\pm$ **0.009**	**0.801** $\pm$ **0.004**	**0.700** $\pm$ **0.007**	**0.736** $\pm$ **0.012**	**0.770** $\pm$ **0.005**
Phi4-mm									
Linear	$\mathcal{O}(1)$	**0.813** $\pm$ **0.004**	**0.700** $\pm$ **0.012**	0.367 $\pm$ 0.019	0.543 $\pm$ 0.015	0.508 $\pm$ 0.008	0.500 $\pm$ 0.015	0.533 $\pm$ 0.015	0.566 $\pm$ 0.009
	$\mathcal{O}(1)$ **- Ours**	0.809 $\pm$ 0.009	0.663 $\pm$ 0.016	**0.696** $\pm$ **0.013**	**0.574** $\pm$ **0.009**	**0.600** $\pm$ **0.010**	**0.559** $\pm$ **0.009**	**0.569** $\pm$ **0.009**	**0.639** $\pm$ **0.010**
	$\mathcal{O}(1)$ **- Updated**	0.868 $\pm$ 0.003	**0.772** $\pm$ **0.004**	0.416 $\pm$ 0.020	0.595 $\pm$ 0.014	0.591 $\pm$ 0.009	0.555 $\pm$ 0.015	0.588 $\pm$ 0.014	0.627 $\pm$ 0.009
	$\mathcal{O}(1)$ **- Ours - Updated**	**0.871** $\pm$ **0.004**	0.771 $\pm$ 0.011	**0.767** $\pm$ **0.009**	**0.651** $\pm$ **0.010**	**0.706** $\pm$ **0.010**	**0.641** $\pm$ **0.011**	**0.649** $\pm$ **0.010**	**0.722** $\pm$ **0.009**
MLP	$\mathcal{O}(1)$	0.820 $\pm$ 0.010	**0.712** $\pm$ **0.018**	0.461 $\pm$ 0.056	0.567 $\pm$ 0.027	0.562 $\pm$ 0.022	0.547 $\pm$ 0.033	0.566 $\pm$ 0.027	0.605 $\pm$ 0.025
	$\mathcal{O}(1)$ **- Ours**	**0.845** $\pm$ **0.003**	0.699 $\pm$ 0.010	**0.684** $\pm$ **0.006**	**0.612** $\pm$ **0.012**	**0.649** $\pm$ **0.005**	**0.590** $\pm$ **0.013**	**0.605** $\pm$ **0.012**	**0.669** $\pm$ **0.007**
	$\mathcal{O}(1)$ **- Updated**	0.889 $\pm$ 0.003	**0.790** $\pm$ **0.010**	0.461 $\pm$ 0.024	0.592 $\pm$ 0.016	0.648 $\pm$ 0.006	0.574 $\pm$ 0.022	0.594 $\pm$ 0.018	0.650 $\pm$ 0.008
	$\mathcal{O}(1)$ **- Ours - Updated**	**0.895** $\pm$ **0.001**	0.770 $\pm$ 0.004	**0.733** $\pm$ **0.013**	**0.658** $\pm$ **0.012**	**0.738** $\pm$ **0.005**	**0.640** $\pm$ **0.012**	**0.654** $\pm$ **0.011**	**0.727** $\pm$ **0.005**
Interactive-Omni-8B									
Linear	$\mathcal{O}(1)$	**0.834** $\pm$ **0.003**	0.835 $\pm$ 0.005	0.542 $\pm$ 0.016	0.694 $\pm$ 0.012	0.619 $\pm$ 0.013	0.643 $\pm$ 0.004	0.656 $\pm$ 0.010	0.689 $\pm$ 0.008
	$\mathcal{O}(1)$ **- Ours**	**0.834** $\pm$ **0.005**	**0.846** $\pm$ **0.007**	**0.724** $\pm$ **0.018**	**0.767** $\pm$ **0.011**	**0.727** $\pm$ **0.010**	**0.734** $\pm$ **0.016**	**0.730** $\pm$ **0.013**	**0.766** $\pm$ **0.011**
	$\mathcal{O}(1)$ **- Updated**	**0.887** $\pm$ **0.002**	0.864 $\pm$ 0.005	0.612 $\pm$ 0.010	0.743 $\pm$ 0.017	0.688 $\pm$ 0.012	0.721 $\pm$ 0.003	0.720 $\pm$ 0.011	0.748 $\pm$ 0.008
	$\mathcal{O}(1)$ **- Ours - Updated**	0.886 $\pm$ 0.003	**0.871** $\pm$ **0.002**	**0.755** $\pm$ **0.013**	**0.810** $\pm$ **0.003**	**0.802** $\pm$ **0.006**	**0.776** $\pm$ **0.005**	**0.786** $\pm$ **0.004**	**0.812** $\pm$ **0.002**
MLP	$\mathcal{O}(1)$	0.839 $\pm$ 0.009	0.852 $\pm$ 0.005	0.578 $\pm$ 0.045	0.710 $\pm$ 0.023	0.648 $\pm$ 0.018	0.646 $\pm$ 0.026	0.666 $\pm$ 0.011	0.705 $\pm$ 0.011
	$\mathcal{O}(1)$ **- Ours**	**0.860** $\pm$ **0.002**	**0.857** $\pm$ **0.005**	**0.736** $\pm$ **0.037**	**0.784** $\pm$ **0.015**	**0.760** $\pm$ **0.024**	**0.735** $\pm$ **0.025**	**0.754** $\pm$ **0.014**	**0.783** $\pm$ **0.017**
	$\mathcal{O}(1)$ **- Updated**	**0.902** $\pm$ **0.002**	**0.883** $\pm$ **0.003**	0.632 $\pm$ 0.025	0.764 $\pm$ 0.017	0.724 $\pm$ 0.008	0.728 $\pm$ 0.015	0.739 $\pm$ 0.008	0.767 $\pm$ 0.006
	$\mathcal{O}(1)$ **- Ours - Updated**	0.894 $\pm$ 0.005	0.875 $\pm$ 0.002	**0.760** $\pm$ **0.025**	**0.812** $\pm$ **0.008**	**0.814** $\pm$ **0.009**	**0.764** $\pm$ **0.007**	**0.798** $\pm$ **0.004**	**0.817** $\pm$ **0.006**

setting. This suggests that the gains are not specific to NoRobots and that modality variation can be estimated effectively from different general, non-safety-related datasets.

Case Study: Sensitivity to hyperparameters. We study the sensitivity of the MLP probe to the subspace dimension k, covariance regularization ρ, and safety-conditioning weight α on WGM-Omni using `Qwen2.5-Omni-3B` and `Phi4-mm`. Figure 9 reports seven-modality AUC for $k \in \{32, 64, 128, 256\}$ over a logarithmic grid of α and ρ, using a fixed seed (42). Performance is largely insensitive to α once $\alpha \geq 1$. In contrast, ρ has a clearer effect, with excessive regularization ($\rho = 10$) consistently reducing AUC, particularly for Phi4-mm; $k = 64$ and $k = 128$ are also generally stronger than the two boundary settings. A stable high-performing region shared by both models is $\alpha \in \{1, 10\}$, $\rho \in \{0.01, 0.1\}$, and $k \in \{64, 128\}$.

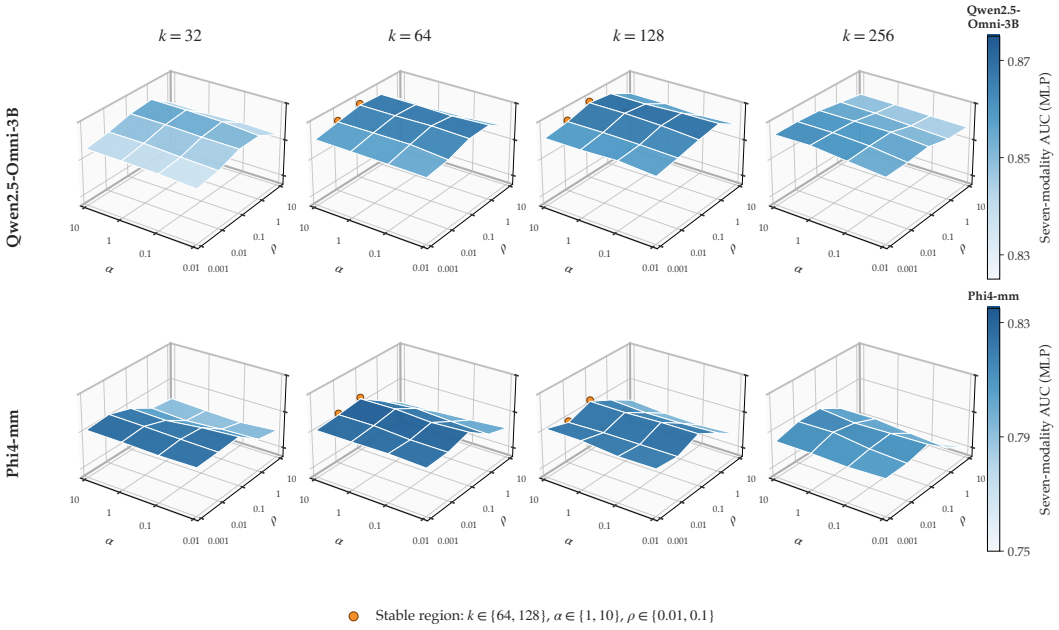


Figure 9: **Sensitivity of MLP probes to α, ρ, and k on WGM-Omni.** Each 3D response surface reports seven-modality AUC for a fixed seed (42); columns vary the subspace dimension k, the horizontal axes vary α and ρ, and both surface height and color encode AUC. Separate color scales are shared across the four k settings within each model row.

Table 26: **Sensitivity of the $\mathcal{O}(1)$ subspace method to the choice of general dataset on Aegis-Omni.** Entries report mean AUC $\pm$ sample standard deviation over five seeds, rounded to three decimals.

Monitor	Method ↓ / Test →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	$\mathcal{O}(1)$ **- Original**	0.901 ± 0.001	0.908 ± 0.002	0.522 ± 0.014	0.745 ± 0.007	0.728 ± 0.017	0.687 ± 0.013	0.738 ± 0.009	0.747 ± 0.007
	$\mathcal{O}(1)$ **- NoRobots**	0.885 ± 0.005	0.882 ± 0.005	0.713 ± 0.010	0.795 ± 0.005	0.824 ± 0.006	0.780 ± 0.003	0.795 ± 0.005	0.810 ± 0.002
	$\mathcal{O}(1)$ **- Dolly**	0.893 ± 0.001	0.895 ± 0.001	0.735 ± 0.005	0.790 ± 0.002	0.836 ± 0.001	0.779 ± 0.006	0.787 ± 0.002	0.816 ± 0.001
MLP	$\mathcal{O}(1)$ **- Original**	0.903 ± 0.003	0.905 ± 0.004	0.536 ± 0.053	0.752 ± 0.015	0.721 ± 0.024	0.729 ± 0.013	0.746 ± 0.015	0.756 ± 0.009
	$\mathcal{O}(1)$ **- NoRobots**	0.901 ± 0.004	0.891 ± 0.005	0.675 ± 0.025	0.794 ± 0.008	0.827 ± 0.008	0.781 ± 0.006	0.794 ± 0.007	0.809 ± 0.006
	$\mathcal{O}(1)$ **- Dolly**	0.907 ± 0.002	0.902 ± 0.001	0.667 ± 0.040	0.763 ± 0.015	0.825 ± 0.009	0.764 ± 0.008	0.762 ± 0.013	0.799 ± 0.011
Phi4-mm									
Linear	$\mathcal{O}(1)$ **- Original**	0.904 ± 0.002	0.850 ± 0.006	0.626 ± 0.039	0.688 ± 0.005	0.760 ± 0.014	0.678 ± 0.008	0.693 ± 0.005	0.743 ± 0.005
	$\mathcal{O}(1)$ **- NoRobots**	0.899 ± 0.002	0.864 ± 0.002	0.818 ± 0.003	0.753 ± 0.005	0.828 ± 0.002	0.760 ± 0.004	0.759 ± 0.004	0.812 ± 0.003
	$\mathcal{O}(1)$ **- Dolly**	0.902 ± 0.001	0.868 ± 0.001	0.808 ± 0.005	0.770 ± 0.002	0.832 ± 0.001	0.772 ± 0.002	0.776 ± 0.002	0.818 ± 0.001
MLP	$\mathcal{O}(1)$ **- Original**	0.903 ± 0.004	0.839 ± 0.011	0.659 ± 0.044	0.686 ± 0.011	0.756 ± 0.022	0.690 ± 0.012	0.694 ± 0.011	0.747 ± 0.008
	$\mathcal{O}(1)$ **- NoRobots**	0.897 ± 0.001	0.844 ± 0.002	0.746 ± 0.016	0.708 ± 0.012	0.782 ± 0.003	0.713 ± 0.010	0.711 ± 0.011	0.771 ± 0.006
	$\mathcal{O}(1)$ **- Dolly**	0.898 ± 0.002	0.844 ± 0.002	0.739 ± 0.017	0.702 ± 0.018	0.795 ± 0.003	0.695 ± 0.013	0.707 ± 0.016	0.769 ± 0.009

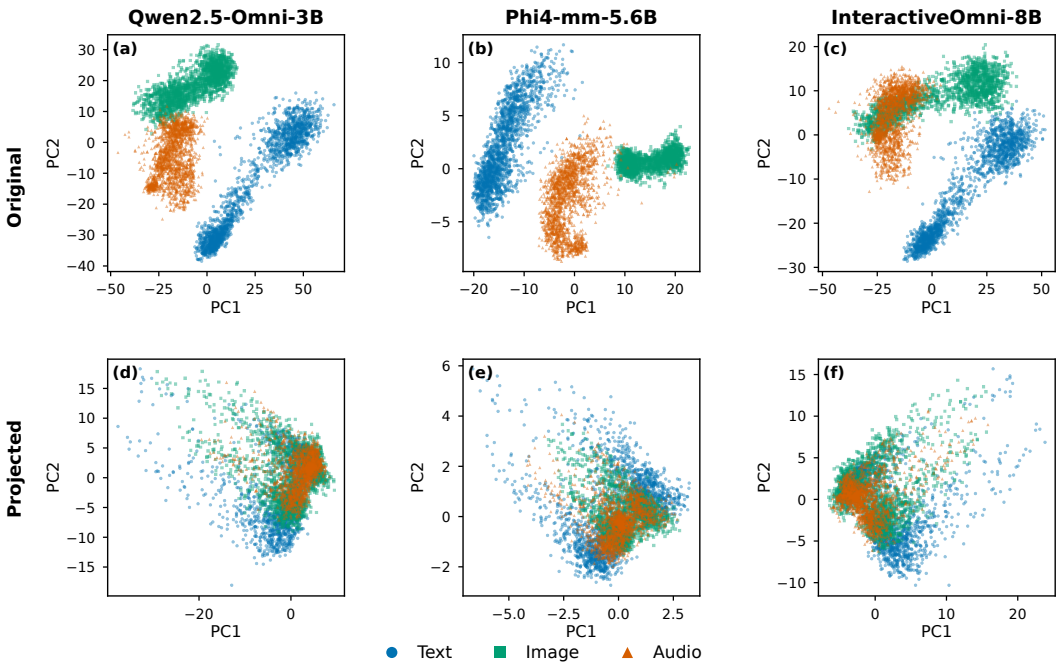


Figure 10: **PCA visualization of original and projected representations on WGM-Omni.** The top row shows the original representations, and the bottom row shows the representations after subspace projection for `Qwen2.5-Omni-3B`, `Phi4-mm-5.6B`, and `InteractiveOmni-8B`. The projected representations exhibit substantially less separation by modality than the original representations.

Table 27: **Mean AUC for the three supervision regimes on WGM-Omni using representations extracted at the $\frac{1}{2}$ layer position.** Entries report mean ± sample standard deviation across five seeds, rounded to three decimals.

Monitor	Method ↓ / Test →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	$\mathcal{O}(1)$	0.930 ± 0.001	0.919 ± 0.005	0.476 ± 0.023	0.788 ± 0.005	0.864 ± 0.004	0.765 ± 0.013	0.772 ± 0.007	0.788 ± 0.006
	$\mathcal{O}(1)$ **- Ours**	0.899 ± 0.001	0.907 ± 0.001	0.788 ± 0.002	0.770 ± 0.001	0.861 ± 0.001	0.812 ± 0.001	0.774 ± 0.001	0.830 ± 0.001
	$\mathcal{O}(n)$	0.928 ± 0.001	0.927 ± 0.003	0.928 ± 0.001	0.797 ± 0.004	0.901 ± 0.002	0.878 ± 0.002	0.809 ± 0.004	0.881 ± 0.002
	$\mathcal{O}(n+1)$	0.924 ± 0.001	0.928 ± 0.002	0.928 ± 0.001	0.859 ± 0.003	0.902 ± 0.002	0.896 ± 0.003	0.869 ± 0.003	0.901 ± 0.001
	$\mathcal{O}(2^n)$	0.930 ± 0.001	0.932 ± 0.001	0.930 ± 0.000	0.873 ± 0.002	0.914 ± 0.002	0.916 ± 0.002	0.875 ± 0.001	0.910 ± 0.001
MLP	$\mathcal{O}(1)$	0.936 ± 0.003	0.921 ± 0.002	0.565 ± 0.030	0.797 ± 0.006	0.887 ± 0.004	0.768 ± 0.015	0.783 ± 0.007	0.808 ± 0.006
	$\mathcal{O}(1)$ **- Ours**	0.916 ± 0.001	0.906 ± 0.003	0.806 ± 0.009	0.764 ± 0.003	0.890 ± 0.002	0.804 ± 0.008	0.771 ± 0.003	0.837 ± 0.003
	$\mathcal{O}(n)$	0.934 ± 0.002	0.925 ± 0.003	0.928 ± 0.002	0.794 ± 0.004	0.908 ± 0.001	0.842 ± 0.009	0.806 ± 0.005	0.877 ± 0.003
	$\mathcal{O}(n+1)$	0.934 ± 0.005	0.929 ± 0.003	0.930 ± 0.003	0.879 ± 0.005	0.918 ± 0.002	0.891 ± 0.010	0.889 ± 0.004	0.910 ± 0.003
	$\mathcal{O}(2^n)$	0.936 ± 0.003	0.937 ± 0.003	0.930 ± 0.001	0.874 ± 0.004	0.918 ± 0.005	0.919 ± 0.002	0.880 ± 0.004	0.913 ± 0.001
Phi4-mm									
Linear	$\mathcal{O}(1)$	0.947 ± 0.002	0.896 ± 0.007	0.566 ± 0.056	0.731 ± 0.005	0.876 ± 0.006	0.596 ± 0.023	0.718 ± 0.005	0.761 ± 0.014
	$\mathcal{O}(1)$ **- Ours**	0.941 ± 0.001	0.899 ± 0.001	0.845 ± 0.001	0.750 ± 0.001	0.868 ± 0.001	0.712 ± 0.000	0.750 ± 0.001	0.824 ± 0.000
	$\mathcal{O}(n)$	0.948 ± 0.002	0.930 ± 0.001	0.938 ± 0.002	0.759 ± 0.003	0.901 ± 0.004	0.755 ± 0.002	0.761 ± 0.003	0.856 ± 0.002
	$\mathcal{O}(n+1)$	0.944 ± 0.001	0.927 ± 0.001	0.936 ± 0.001	0.849 ± 0.003	0.905 ± 0.002	0.866 ± 0.003	0.854 ± 0.003	0.897 ± 0.001
	$\mathcal{O}(2^n)$	0.947 ± 0.002	0.927 ± 0.002	0.940 ± 0.001	0.873 ± 0.002	0.920 ± 0.002	0.902 ± 0.002	0.873 ± 0.002	0.912 ± 0.001
MLP	$\mathcal{O}(1)$	0.952 ± 0.001	0.901 ± 0.008	0.624 ± 0.059	0.750 ± 0.009	0.889 ± 0.006	0.628 ± 0.029	0.741 ± 0.010	0.783 ± 0.016
	$\mathcal{O}(1)$ **- Ours**	0.946 ± 0.001	0.909 ± 0.002	0.863 ± 0.005	0.747 ± 0.007	0.900 ± 0.003	0.694 ± 0.004	0.743 ± 0.007	0.829 ± 0.003
	$\mathcal{O}(n)$	0.947 ± 0.004	0.931 ± 0.005	0.935 ± 0.004	0.736 ± 0.009	0.904 ± 0.004	0.749 ± 0.012	0.739 ± 0.010	0.849 ± 0.005
	$\mathcal{O}(n+1)$	0.947 ± 0.001	0.932 ± 0.003	0.936 ± 0.003	0.863 ± 0.005	0.910 ± 0.002	0.882 ± 0.006	0.867 ± 0.005	0.905 ± 0.002
	$\mathcal{O}(2^n)$	0.952 ± 0.001	0.932 ± 0.004	0.940 ± 0.001	0.870 ± 0.003	0.924 ± 0.002	0.905 ± 0.002	0.872 ± 0.006	0.913 ± 0.001

Table 28: **Mean AUC for the three supervision regimes on WGM-Omni using representations extracted at the $\frac{3}{4}$ layer position.** Entries report mean ± sample standard deviation across five seeds, rounded to three decimals.

Monitor	Method ↓ / Test →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	$\mathcal{O}(1)$	0.938 ± 0.001	0.940 ± 0.002	0.633 ± 0.012	0.817 ± 0.003	0.885 ± 0.002	0.719 ± 0.012	0.805 ± 0.004	0.820 ± 0.003
	$\mathcal{O}(1)$ **- Ours**	0.917 ± 0.001	0.936 ± 0.001	0.821 ± 0.002	0.814 ± 0.001	0.888 ± 0.001	0.809 ± 0.001	0.819 ± 0.001	0.858 ± 0.001
	$\mathcal{O}(n)$	0.933 ± 0.001	0.938 ± 0.001	0.936 ± 0.001	0.821 ± 0.004	0.910 ± 0.001	0.879 ± 0.005	0.831 ± 0.004	0.892 ± 0.002
	$\mathcal{O}(n+1)$	0.931 ± 0.001	0.936 ± 0.001	0.934 ± 0.001	0.878 ± 0.001	0.914 ± 0.001	0.900 ± 0.005	0.886 ± 0.001	0.911 ± 0.001
	$\mathcal{O}(2^n)$	0.938 ± 0.001	0.936 ± 0.002	0.933 ± 0.001	0.883 ± 0.002	0.921 ± 0.001	0.921 ± 0.001	0.884 ± 0.002	0.917 ± 0.001
MLP	$\mathcal{O}(1)$	0.942 ± 0.004	0.942 ± 0.001	0.646 ± 0.025	0.820 ± 0.006	0.893 ± 0.007	0.710 ± 0.010	0.809 ± 0.008	0.823 ± 0.007
	$\mathcal{O}(1)$ **- Ours**	0.926 ± 0.001	0.931 ± 0.002	0.859 ± 0.007	0.816 ± 0.003	0.907 ± 0.002	0.832 ± 0.003	0.824 ± 0.003	0.871 ± 0.001
	$\mathcal{O}(n)$	0.939 ± 0.001	0.937 ± 0.003	0.932 ± 0.003	0.795 ± 0.008	0.917 ± 0.002	0.845 ± 0.011	0.805 ± 0.009	0.881 ± 0.004
	$\mathcal{O}(n+1)$	0.940 ± 0.001	0.937 ± 0.003	0.935 ± 0.002	0.891 ± 0.002	0.926 ± 0.002	0.893 ± 0.007	0.899 ± 0.002	0.917 ± 0.002
	$\mathcal{O}(2^n)$	0.942 ± 0.004	0.940 ± 0.002	0.935 ± 0.002	0.885 ± 0.004	0.925 ± 0.003	0.925 ± 0.002	0.887 ± 0.003	0.920 ± 0.001
Phi4-mm									
Linear	$\mathcal{O}(1)$	0.943 ± 0.002	0.880 ± 0.006	0.595 ± 0.012	0.745 ± 0.003	0.883 ± 0.002	0.651 ± 0.009	0.736 ± 0.004	0.776 ± 0.003
	$\mathcal{O}(1)$ **- Ours**	0.931 ± 0.000	0.879 ± 0.001	0.858 ± 0.001	0.764 ± 0.001	0.876 ± 0.001	0.709 ± 0.002	0.766 ± 0.001	0.826 ± 0.001
	$\mathcal{O}(n)$	0.939 ± 0.001	0.925 ± 0.002	0.935 ± 0.002	0.780 ± 0.004	0.894 ± 0.003	0.773 ± 0.013	0.781 ± 0.005	0.861 ± 0.004
	$\mathcal{O}(n+1)$	0.939 ± 0.001	0.928 ± 0.001	0.936 ± 0.001	0.855 ± 0.002	0.902 ± 0.001	0.873 ± 0.003	0.858 ± 0.002	0.899 ± 0.001
	$\mathcal{O}(2^n)$	0.943 ± 0.002	0.925 ± 0.002	0.938 ± 0.002	0.871 ± 0.002	0.914 ± 0.002	0.904 ± 0.001	0.873 ± 0.002	0.910 ± 0.001
MLP	$\mathcal{O}(1)$	0.945 ± 0.002	0.885 ± 0.003	0.682 ± 0.043	0.757 ± 0.008	0.895 ± 0.004	0.678 ± 0.018	0.752 ± 0.008	0.799 ± 0.009
	$\mathcal{O}(1)$ **- Ours**	0.940 ± 0.001	0.892 ± 0.003	0.866 ± 0.003	0.760 ± 0.008	0.900 ± 0.002	0.719 ± 0.006	0.762 ± 0.009	0.834 ± 0.004
	$\mathcal{O}(n)$	0.941 ± 0.003	0.930 ± 0.003	0.933 ± 0.002	0.773 ± 0.010	0.900 ± 0.006	0.779 ± 0.010	0.774 ± 0.010	0.861 ± 0.004
	$\mathcal{O}(n+1)$	0.941 ± 0.003	0.930 ± 0.004	0.934 ± 0.002	0.874 ± 0.004	0.907 ± 0.005	0.892 ± 0.002	0.878 ± 0.003	0.908 ± 0.001
	$\mathcal{O}(2^n)$	0.945 ± 0.002	0.930 ± 0.002	0.939 ± 0.002	0.870 ± 0.006	0.919 ± 0.003	0.908 ± 0.001	0.875 ± 0.005	0.912 ± 0.001

Table 29: **Mean AUC for the three supervision regimes on WGM-Omni.** Entries report mean ± sample standard deviation across five seeds, rounded to three decimals.

Monitor	Method ↓ / Test →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	$\mathcal{O}(1)$	0.930 ± 0.002	0.927 ± 0.002	0.667 ± 0.023	0.790 ± 0.008	0.871 ± 0.003	0.714 ± 0.020	0.785 ± 0.008	0.812 ± 0.007
	$\mathcal{O}(1)$ **- Ours**	0.923 ± 0.006	0.932 ± 0.001	0.802 ± 0.006	0.801 ± 0.004	0.887 ± 0.003	0.816 ± 0.005	0.803 ± 0.005	0.852 ± 0.002
	$\mathcal{O}(n)$	0.926 ± 0.001	0.931 ± 0.001	0.928 ± 0.001	0.809 ± 0.004	0.884 ± 0.001	0.893 ± 0.004	0.818 ± 0.004	0.884 ± 0.002
	$\mathcal{O}(n+1)$	0.921 ± 0.001	0.930 ± 0.001	0.925 ± 0.001	0.862 ± 0.002	0.893 ± 0.002	0.901 ± 0.003	0.870 ± 0.002	0.900 ± 0.001
	$\mathcal{O}(2^n)$	0.930 ± 0.002	0.933 ± 0.003	0.929 ± 0.002	0.864 ± 0.002	0.903 ± 0.001	0.909 ± 0.002	0.870 ± 0.001	0.906 ± 0.001
MLP	$\mathcal{O}(1)$	0.935 ± 0.002	0.922 ± 0.008	0.688 ± 0.017	0.775 ± 0.013	0.879 ± 0.003	0.707 ± 0.026	0.767 ± 0.010	0.810 ± 0.009
	$\mathcal{O}(1)$ **- Ours**	0.933 ± 0.003	0.934 ± 0.001	0.844 ± 0.012	0.811 ± 0.007	0.901 ± 0.002	0.844 ± 0.007	0.814 ± 0.008	0.869 ± 0.001
	$\mathcal{O}(n)$	0.935 ± 0.001	0.935 ± 0.003	0.931 ± 0.004	0.800 ± 0.005	0.890 ± 0.002	0.876 ± 0.010	0.809 ± 0.005	0.882 ± 0.003
	$\mathcal{O}(n+1)$	0.934 ± 0.003	0.935 ± 0.002	0.932 ± 0.001	0.874 ± 0.002	0.904 ± 0.005	0.906 ± 0.003	0.884 ± 0.001	0.910 ± 0.001
	$\mathcal{O}(2^n)$	0.935 ± 0.002	0.936 ± 0.003	0.933 ± 0.002	0.869 ± 0.002	0.912 ± 0.004	0.912 ± 0.001	0.877 ± 0.003	0.911 ± 0.001
Qwen2.5-Omni-7B									
Linear	$\mathcal{O}(1)$	0.933 ± 0.003	0.943 ± 0.002	0.911 ± 0.005	0.818 ± 0.007	0.889 ± 0.003	0.801 ± 0.016	0.825 ± 0.007	0.874 ± 0.005
	$\mathcal{O}(1)$ **- Ours**	0.924 ± 0.001	0.936 ± 0.002	0.889 ± 0.007	0.824 ± 0.002	0.888 ± 0.003	0.831 ± 0.009	0.824 ± 0.001	0.874 ± 0.003
	$\mathcal{O}(n)$	0.936 ± 0.001	0.943 ± 0.001	0.924 ± 0.002	0.809 ± 0.004	0.890 ± 0.003	0.890 ± 0.003	0.819 ± 0.003	0.887 ± 0.002
	$\mathcal{O}(n+1)$	0.933 ± 0.002	0.942 ± 0.001	0.926 ± 0.003	0.866 ± 0.003	0.911 ± 0.001	0.913 ± 0.003	0.871 ± 0.003	0.909 ± 0.002
	$\mathcal{O}(2^n)$	0.933 ± 0.003	0.944 ± 0.001	0.911 ± 0.005	0.882 ± 0.002	0.917 ± 0.001	0.919 ± 0.000	0.884 ± 0.002	0.913 ± 0.001
MLP	$\mathcal{O}(1)$	0.937 ± 0.003	0.939 ± 0.002	0.901 ± 0.006	0.804 ± 0.012	0.895 ± 0.003	0.791 ± 0.025	0.806 ± 0.012	0.868 ± 0.007
	$\mathcal{O}(1)$ **- Ours**	0.932 ± 0.004	0.938 ± 0.002	0.904 ± 0.005	0.816 ± 0.005	0.899 ± 0.006	0.846 ± 0.012	0.819 ± 0.005	0.879 ± 0.005
	$\mathcal{O}(n)$	0.941 ± 0.002	0.945 ± 0.001	0.927 ± 0.002	0.819 ± 0.005	0.902 ± 0.003	0.897 ± 0.005	0.830 ± 0.004	0.894 ± 0.002
	$\mathcal{O}(n+1)$	0.938 ± 0.002	0.944 ± 0.002	0.931 ± 0.001	0.883 ± 0.005	0.914 ± 0.002	0.919 ± 0.003	0.889 ± 0.005	0.917 ± 0.002
	$\mathcal{O}(2^n)$	0.937 ± 0.003	0.946 ± 0.002	0.918 ± 0.003	0.881 ± 0.004	0.918 ± 0.003	0.923 ± 0.002	0.887 ± 0.002	0.916 ± 0.001
Qwen3-Omni-30B									
Linear	$\mathcal{O}(1)$	0.942 ± 0.001	0.903 ± 0.016	0.669 ± 0.031	0.841 ± 0.011	0.910 ± 0.002	0.577 ± 0.032	0.813 ± 0.010	0.808 ± 0.013
	$\mathcal{O}(1)$ **- Ours**	0.935 ± 0.001	0.918 ± 0.000	0.833 ± 0.014	0.864 ± 0.001	0.911 ± 0.002	0.782 ± 0.008	0.847 ± 0.001	0.870 ± 0.002
	$\mathcal{O}(n)$	0.936 ± 0.001	0.928 ± 0.002	0.935 ± 0.001	0.857 ± 0.002	0.916 ± 0.001	0.875 ± 0.004	0.850 ± 0.001	0.899 ± 0.001
	$\mathcal{O}(n+1)$	0.936 ± 0.001	0.931 ± 0.001	0.934 ± 0.001	0.889 ± 0.001	0.921 ± 0.001	0.891 ± 0.002	0.883 ± 0.001	0.912 ± 0.001
	$\mathcal{O}(2^n)$	0.942 ± 0.001	0.929 ± 0.002	0.934 ± 0.001	0.902 ± 0.001	0.926 ± 0.001	0.908 ± 0.001	0.889 ± 0.002	0.918 ± 0.001
MLP	$\mathcal{O}(1)$	0.943 ± 0.003	0.916 ± 0.005	0.709 ± 0.047	0.831 ± 0.022	0.917 ± 0.005	0.622 ± 0.046	0.808 ± 0.017	0.821 ± 0.019
	$\mathcal{O}(1)$ **- Ours**	0.937 ± 0.002	0.912 ± 0.005	0.860 ± 0.005	0.846 ± 0.006	0.917 ± 0.001	0.785 ± 0.014	0.828 ± 0.005	0.869 ± 0.002
	$\mathcal{O}(n)$	0.941 ± 0.002	0.930 ± 0.006	0.938 ± 0.002	0.854 ± 0.010	0.920 ± 0.002	0.829 ± 0.008	0.846 ± 0.007	0.894 ± 0.003
	$\mathcal{O}(n+1)$	0.940 ± 0.002	0.933 ± 0.001	0.938 ± 0.002	0.901 ± 0.002	0.922 ± 0.002	0.886 ± 0.007	0.895 ± 0.003	0.917 ± 0.001
	$\mathcal{O}(2^n)$	0.943 ± 0.003	0.932 ± 0.003	0.938 ± 0.002	0.908 ± 0.002	0.929 ± 0.002	0.913 ± 0.002	0.895 ± 0.002	0.923 ± 0.001
Phi4-mm 5.6B									
Linear	$\mathcal{O}(1)$	0.940 ± 0.002	0.883 ± 0.009	0.637 ± 0.034	0.707 ± 0.006	0.834 ± 0.002	0.605 ± 0.011	0.701 ± 0.006	0.758 ± 0.007
	$\mathcal{O}(1)$ **- Ours**	0.931 ± 0.004	0.863 ± 0.011	0.845 ± 0.006	0.725 ± 0.012	0.827 ± 0.007	0.671 ± 0.012	0.724 ± 0.012	0.798 ± 0.009
	$\mathcal{O}(n)$	0.941 ± 0.001	0.922 ± 0.003	0.934 ± 0.001	0.756 ± 0.003	0.856 ± 0.004	0.745 ± 0.008	0.761 ± 0.003	0.845 ± 0.003
	$\mathcal{O}(n+1)$	0.939 ± 0.001	0.920 ± 0.002	0.936 ± 0.001	0.833 ± 0.002	0.867 ± 0.005	0.848 ± 0.002	0.838 ± 0.002	0.883 ± 0.001
	$\mathcal{O}(2^n)$	0.940 ± 0.002	0.922 ± 0.002	0.935 ± 0.001	0.854 ± 0.001	0.899 ± 0.003	0.899 ± 0.002	0.858 ± 0.001	0.901 ± 0.000
MLP	$\mathcal{O}(1)$	0.944 ± 0.002	0.874 ± 0.023	0.780 ± 0.044	0.709 ± 0.012	0.844 ± 0.009	0.618 ± 0.034	0.708 ± 0.011	0.782 ± 0.015
	$\mathcal{O}(1)$ **- Ours**	0.943 ± 0.001	0.895 ± 0.006	0.860 ± 0.005	0.742 ± 0.008	0.877 ± 0.004	0.706 ± 0.015	0.740 ± 0.009	0.823 ± 0.006
	$\mathcal{O}(n)$	0.943 ± 0.002	0.923 ± 0.007	0.935 ± 0.002	0.741 ± 0.009	0.862 ± 0.007	0.753 ± 0.015	0.746 ± 0.009	0.843 ± 0.005
	$\mathcal{O}(n+1)$	0.943 ± 0.001	0.925 ± 0.003	0.934 ± 0.002	0.850 ± 0.004	0.860 ± 0.011	0.867 ± 0.005	0.855 ± 0.004	0.891 ± 0.002
	$\mathcal{O}(2^n)$	0.944 ± 0.002	0.924 ± 0.003	0.935 ± 0.003	0.850 ± 0.003	0.896 ± 0.003	0.898 ± 0.002	0.856 ± 0.003	0.900 ± 0.000
Interactive-Omni-8B									
Linear	$\mathcal{O}(1)$	0.941 ± 0.001	0.907 ± 0.010	0.858 ± 0.010	0.823 ± 0.009	0.879 ± 0.005	0.842 ± 0.012	0.835 ± 0.004	0.869 ± 0.007
	$\mathcal{O}(1)$ **- Ours**	0.929 ± 0.005	0.926 ± 0.002	0.858 ± 0.012	0.852 ± 0.003	0.882 ± 0.008	0.851 ± 0.006	0.856 ± 0.004	0.879 ± 0.006
	$\mathcal{O}(n)$	0.937 ± 0.002	0.930 ± 0.001	0.932 ± 0.001	0.851 ± 0.004	0.902 ± 0.002	0.911 ± 0.001	0.875 ± 0.002	0.905 ± 0.002
	$\mathcal{O}(n+1)$	0.933 ± 0.002	0.926 ± 0.001	0.927 ± 0.002	0.880 ± 0.003	0.908 ± 0.002	0.910 ± 0.003	0.898 ± 0.002	0.912 ± 0.002
	$\mathcal{O}(2^n)$	0.941 ± 0.001	0.930 ± 0.001	0.929 ± 0.003	0.904 ± 0.002	0.915 ± 0.002	0.916 ± 0.001	0.905 ± 0.001	0.920 ± 0.001
MLP	$\mathcal{O}(1)$	0.941 ± 0.002	0.922 ± 0.007	0.860 ± 0.017	0.854 ± 0.008	0.891 ± 0.004	0.852 ± 0.012	0.850 ± 0.002	0.881 ± 0.003
	$\mathcal{O}(1)$ **- Ours**	0.937 ± 0.002	0.929 ± 0.003	0.874 ± 0.005	0.853 ± 0.009	0.895 ± 0.005	0.860 ± 0.005	0.865 ± 0.007	0.887 ± 0.005
	$\mathcal{O}(n)$	0.939 ± 0.004	0.931 ± 0.002	0.930 ± 0.004	0.849 ± 0.007	0.903 ± 0.004	0.910 ± 0.005	0.876 ± 0.002	0.905 ± 0.002
	$\mathcal{O}(n+1)$	0.940 ± 0.002	0.928 ± 0.003	0.927 ± 0.005	0.885 ± 0.002	0.914 ± 0.002	0.910 ± 0.007	0.908 ± 0.004	0.916 ± 0.003
	$\mathcal{O}(2^n)$	0.941 ± 0.002	0.926 ± 0.007	0.932 ± 0.002	0.900 ± 0.006	0.915 ± 0.004	0.920 ± 0.003	0.907 ± 0.005	0.920 ± 0.003

Table 30: **Mean accuracy for the three supervision regimes on WGM-Omni.** Entries report mean ± sample standard deviation across five seeds, rounded to three decimals.

Monitor	Method ↓ / Test →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	$\mathcal{O}(1)$	0.861 ± 0.003	0.850 ± 0.019	0.580 ± 0.011	0.703 ± 0.017	0.810 ± 0.007	0.650 ± 0.020	0.699 ± 0.024	0.736 ± 0.006
	$\mathcal{O}(1)$ **- Ours**	0.850 ± 0.006	0.871 ± 0.003	0.681 ± 0.010	0.742 ± 0.007	0.813 ± 0.004	0.699 ± 0.027	0.741 ± 0.012	0.771 ± 0.004
	$\mathcal{O}(n)$	0.860 ± 0.003	0.875 ± 0.003	0.870 ± 0.003	0.728 ± 0.005	0.805 ± 0.003	0.784 ± 0.012	0.737 ± 0.005	0.809 ± 0.003
	$\mathcal{O}(n+1)$	0.857 ± 0.005	0.874 ± 0.001	0.862 ± 0.002	0.788 ± 0.004	0.817 ± 0.005	0.790 ± 0.017	0.805 ± 0.002	0.828 ± 0.002
	$\mathcal{O}(2^n)$	0.861 ± 0.003	0.873 ± 0.003	0.868 ± 0.003	0.799 ± 0.002	0.840 ± 0.005	0.844 ± 0.003	0.812 ± 0.002	0.842 ± 0.001
MLP	$\mathcal{O}(1)$	0.866 ± 0.003	0.842 ± 0.024	0.596 ± 0.015	0.697 ± 0.021	0.813 ± 0.002	0.648 ± 0.020	0.693 ± 0.016	0.736 ± 0.010
	$\mathcal{O}(1)$ **- Ours**	0.865 ± 0.004	0.876 ± 0.004	0.696 ± 0.010	0.721 ± 0.031	0.830 ± 0.007	0.749 ± 0.006	0.725 ± 0.033	0.780 ± 0.008
	$\mathcal{O}(n)$	0.869 ± 0.004	0.878 ± 0.002	0.871 ± 0.005	0.704 ± 0.015	0.814 ± 0.006	0.750 ± 0.031	0.720 ± 0.015	0.801 ± 0.008
	$\mathcal{O}(n+1)$	0.869 ± 0.003	0.882 ± 0.004	0.865 ± 0.005	0.795 ± 0.003	0.832 ± 0.007	0.818 ± 0.018	0.810 ± 0.008	0.839 ± 0.005
	$\mathcal{O}(2^n)$	0.866 ± 0.003	0.877 ± 0.004	0.867 ± 0.005	0.799 ± 0.004	0.846 ± 0.008	0.844 ± 0.007	0.805 ± 0.006	0.843 ± 0.002
Qwen2.5-Omni-7B									
Linear	$\mathcal{O}(1)$	0.873 ± 0.008	0.873 ± 0.011	0.713 ± 0.039	0.744 ± 0.009	0.740 ± 0.005	0.654 ± 0.025	0.739 ± 0.016	0.762 ± 0.014
	$\mathcal{O}(1)$ **- Ours**	0.848 ± 0.002	0.862 ± 0.004	0.787 ± 0.010	0.735 ± 0.001	0.811 ± 0.004	0.750 ± 0.009	0.745 ± 0.002	0.791 ± 0.004
	$\mathcal{O}(n)$	0.879 ± 0.003	0.887 ± 0.003	0.828 ± 0.009	0.680 ± 0.006	0.821 ± 0.004	0.799 ± 0.006	0.692 ± 0.005	0.798 ± 0.003
	$\mathcal{O}(n+1)$	0.870 ± 0.007	0.886 ± 0.001	0.835 ± 0.009	0.798 ± 0.006	0.846 ± 0.002	0.853 ± 0.005	0.807 ± 0.007	0.842 ± 0.002
	$\mathcal{O}(2^n)$	0.873 ± 0.008	0.887 ± 0.003	0.800 ± 0.006	0.810 ± 0.005	0.852 ± 0.001	0.854 ± 0.002	0.816 ± 0.005	0.842 ± 0.003
MLP	$\mathcal{O}(1)$	0.877 ± 0.006	0.869 ± 0.009	0.774 ± 0.043	0.701 ± 0.019	0.794 ± 0.019	0.718 ± 0.025	0.715 ± 0.019	0.778 ± 0.008
	$\mathcal{O}(1)$ **- Ours**	0.870 ± 0.005	0.870 ± 0.005	0.824 ± 0.007	0.742 ± 0.006	0.826 ± 0.007	0.771 ± 0.010	0.745 ± 0.005	0.807 ± 0.005
	$\mathcal{O}(n)$	0.880 ± 0.003	0.892 ± 0.002	0.839 ± 0.005	0.704 ± 0.016	0.833 ± 0.006	0.815 ± 0.002	0.723 ± 0.016	0.812 ± 0.005
	$\mathcal{O}(n+1)$	0.878 ± 0.008	0.891 ± 0.003	0.837 ± 0.006	0.816 ± 0.005	0.852 ± 0.005	0.858 ± 0.005	0.823 ± 0.006	0.851 ± 0.002
	$\mathcal{O}(2^n)$	0.877 ± 0.006	0.886 ± 0.001	0.801 ± 0.013	0.811 ± 0.005	0.853 ± 0.006	0.859 ± 0.004	0.819 ± 0.003	0.844 ± 0.001
Qwen3-Omni-30B									
Linear	$\mathcal{O}(1)$	0.875 ± 0.003	0.727 ± 0.053	0.592 ± 0.021	0.656 ± 0.083	0.848 ± 0.007	0.588 ± 0.014	0.657 ± 0.068	0.706 ± 0.032
	$\mathcal{O}(1)$ **- Ours**	0.865 ± 0.002	0.806 ± 0.025	0.680 ± 0.013	0.686 ± 0.026	0.854 ± 0.001	0.644 ± 0.033	0.692 ± 0.029	0.747 ± 0.015
	$\mathcal{O}(n)$	0.872 ± 0.003	0.860 ± 0.008	0.873 ± 0.002	0.714 ± 0.021	0.856 ± 0.003	0.801 ± 0.009	0.728 ± 0.022	0.815 ± 0.008
	$\mathcal{O}(n+1)$	0.868 ± 0.001	0.869 ± 0.005	0.874 ± 0.003	0.826 ± 0.004	0.862 ± 0.002	0.829 ± 0.004	0.816 ± 0.002	0.849 ± 0.001
	$\mathcal{O}(2^n)$	0.875 ± 0.003	0.854 ± 0.003	0.868 ± 0.001	0.845 ± 0.003	0.862 ± 0.002	0.845 ± 0.004	0.829 ± 0.003	0.854 ± 0.001
MLP	$\mathcal{O}(1)$	0.877 ± 0.006	0.666 ± 0.057	0.612 ± 0.027	0.643 ± 0.050	0.847 ± 0.009	0.565 ± 0.034	0.632 ± 0.046	0.692 ± 0.016
	$\mathcal{O}(1)$ **- Ours**	0.875 ± 0.003	0.816 ± 0.032	0.726 ± 0.013	0.722 ± 0.041	0.860 ± 0.005	0.688 ± 0.052	0.718 ± 0.041	0.772 ± 0.023
	$\mathcal{O}(n)$	0.878 ± 0.003	0.862 ± 0.008	0.876 ± 0.005	0.744 ± 0.025	0.856 ± 0.004	0.730 ± 0.018	0.753 ± 0.017	0.814 ± 0.007
	$\mathcal{O}(n+1)$	0.879 ± 0.005	0.874 ± 0.008	0.879 ± 0.005	0.830 ± 0.007	0.863 ± 0.003	0.818 ± 0.011	0.830 ± 0.004	0.853 ± 0.003
	$\mathcal{O}(2^n)$	0.877 ± 0.006	0.866 ± 0.006	0.874 ± 0.004	0.844 ± 0.006	0.864 ± 0.007	0.847 ± 0.003	0.831 ± 0.006	0.857 ± 0.003
Phi-4-multimodal									
Linear	$\mathcal{O}(1)$	0.875 ± 0.003	0.762 ± 0.033	0.586 ± 0.004	0.574 ± 0.009	0.762 ± 0.016	0.569 ± 0.005	0.571 ± 0.007	0.671 ± 0.008
	$\mathcal{O}(1)$ **- Ours**	0.850 ± 0.010	0.706 ± 0.019	0.589 ± 0.004	0.590 ± 0.004	0.696 ± 0.008	0.574 ± 0.004	0.589 ± 0.004	0.656 ± 0.007
	$\mathcal{O}(n)$	0.878 ± 0.005	0.862 ± 0.003	0.872 ± 0.003	0.680 ± 0.011	0.778 ± 0.011	0.679 ± 0.011	0.687 ± 0.011	0.777 ± 0.006
	$\mathcal{O}(n+1)$	0.880 ± 0.003	0.851 ± 0.003	0.874 ± 0.001	0.771 ± 0.003	0.790 ± 0.007	0.768 ± 0.002	0.772 ± 0.001	0.815 ± 0.001
	$\mathcal{O}(2^n)$	0.875 ± 0.003	0.851 ± 0.005	0.869 ± 0.002	0.791 ± 0.003	0.831 ± 0.005	0.836 ± 0.003	0.794 ± 0.004	0.835 ± 0.001
MLP	$\mathcal{O}(1)$	0.879 ± 0.003	0.736 ± 0.047	0.622 ± 0.039	0.566 ± 0.010	0.758 ± 0.015	0.564 ± 0.007	0.567 ± 0.011	0.670 ± 0.014
	$\mathcal{O}(1)$ **- Ours**	0.876 ± 0.004	0.719 ± 0.009	0.603 ± 0.005	0.613 ± 0.012	0.735 ± 0.009	0.597 ± 0.009	0.609 ± 0.012	0.679 ± 0.007
	$\mathcal{O}(n)$	0.882 ± 0.003	0.856 ± 0.008	0.875 ± 0.003	0.652 ± 0.025	0.789 ± 0.004	0.663 ± 0.031	0.660 ± 0.027	0.768 ± 0.014
	$\mathcal{O}(n+1)$	0.887 ± 0.003	0.858 ± 0.003	0.871 ± 0.004	0.769 ± 0.012	0.790 ± 0.009	0.782 ± 0.017	0.784 ± 0.009	0.820 ± 0.004
	$\mathcal{O}(2^n)$	0.879 ± 0.003	0.857 ± 0.004	0.870 ± 0.005	0.780 ± 0.008	0.828 ± 0.006	0.834 ± 0.005	0.787 ± 0.004	0.834 ± 0.002
Interactive-Omni-8B									
Linear	$\mathcal{O}(1)$	0.876 ± 0.002	0.797 ± 0.022	0.577 ± 0.013	0.617 ± 0.020	0.814 ± 0.008	0.578 ± 0.012	0.664 ± 0.032	0.703 ± 0.015
	$\mathcal{O}(1)$ **- Ours**	0.856 ± 0.007	0.856 ± 0.007	0.706 ± 0.015	0.781 ± 0.004	0.813 ± 0.010	0.771 ± 0.008	0.787 ± 0.002	0.796 ± 0.006
	$\mathcal{O}(n)$	0.873 ± 0.003	0.873 ± 0.006	0.871 ± 0.004	0.769 ± 0.004	0.825 ± 0.003	0.845 ± 0.002	0.804 ± 0.004	0.837 ± 0.001
	$\mathcal{O}(n+1)$	0.867 ± 0.005	0.866 ± 0.005	0.868 ± 0.003	0.805 ± 0.003	0.842 ± 0.003	0.846 ± 0.003	0.837 ± 0.005	0.847 ± 0.001
	$\mathcal{O}(2^n)$	0.876 ± 0.002	0.868 ± 0.004	0.868 ± 0.006	0.842 ± 0.002	0.845 ± 0.008	0.851 ± 0.006	0.836 ± 0.002	0.855 ± 0.001
MLP	$\mathcal{O}(1)$	0.878 ± 0.001	0.820 ± 0.023	0.617 ± 0.039	0.687 ± 0.036	0.820 ± 0.011	0.618 ± 0.034	0.725 ± 0.034	0.738 ± 0.022
	$\mathcal{O}(1)$ **- Ours**	0.874 ± 0.003	0.861 ± 0.007	0.764 ± 0.050	0.779 ± 0.005	0.830 ± 0.005	0.778 ± 0.008	0.795 ± 0.008	0.812 ± 0.006
	$\mathcal{O}(n)$	0.876 ± 0.006	0.878 ± 0.003	0.871 ± 0.006	0.764 ± 0.017	0.826 ± 0.004	0.840 ± 0.009	0.807 ± 0.007	0.837 ± 0.002
	$\mathcal{O}(n+1)$	0.881 ± 0.004	0.876 ± 0.004	0.869 ± 0.003	0.808 ± 0.009	0.849 ± 0.003	0.853 ± 0.009	0.846 ± 0.007	0.855 ± 0.003
	$\mathcal{O}(2^n)$	0.878 ± 0.001	0.865 ± 0.004	0.872 ± 0.007	0.835 ± 0.005	0.847 ± 0.011	0.856 ± 0.007	0.839 ± 0.009	0.856 ± 0.005

Table 31: **Mean f1 score for the three supervision regimes on WGM-Omni.** Entries report mean ± sample standard deviation across five seeds, rounded to three decimals.

Monitor	Method ↓ / Test →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	$\mathcal{O}(1)$	0.830 ± 0.003	0.812 ± 0.034	0.377 ± 0.054	0.656 ± 0.054	0.770 ± 0.014	0.494 ± 0.094	0.661 ± 0.044	0.657 ± 0.036
	$\mathcal{O}(1)$ **- Ours**	0.815 ± 0.006	0.850 ± 0.002	0.572 ± 0.032	0.695 ± 0.010	0.781 ± 0.008	0.553 ± 0.082	0.705 ± 0.010	0.710 ± 0.016
	$\mathcal{O}(n)$	0.831 ± 0.004	0.851 ± 0.004	0.849 ± 0.005	0.623 ± 0.014	0.756 ± 0.005	0.709 ± 0.022	0.649 ± 0.011	0.753 ± 0.006
	$\mathcal{O}(n+1)$	0.827 ± 0.007	0.850 ± 0.002	0.840 ± 0.003	0.747 ± 0.009	0.773 ± 0.006	0.719 ± 0.032	0.776 ± 0.005	0.790 ± 0.005
	$\mathcal{O}(2^n)$	0.830 ± 0.003	0.847 ± 0.004	0.845 ± 0.005	0.769 ± 0.003	0.806 ± 0.007	0.817 ± 0.004	0.784 ± 0.003	0.814 ± 0.001
MLP	$\mathcal{O}(1)$	0.838 ± 0.005	0.806 ± 0.045	0.344 ± 0.077	0.638 ± 0.089	0.769 ± 0.008	0.485 ± 0.079	0.637 ± 0.083	0.645 ± 0.035
	$\mathcal{O}(1)$ **- Ours**	0.836 ± 0.005	0.857 ± 0.005	0.573 ± 0.033	0.713 ± 0.015	0.798 ± 0.011	0.674 ± 0.016	0.720 ± 0.011	0.739 ± 0.007
	$\mathcal{O}(n)$	0.843 ± 0.008	0.856 ± 0.002	0.851 ± 0.008	0.579 ± 0.042	0.779 ± 0.012	0.650 ± 0.066	0.624 ± 0.040	0.740 ± 0.018
	$\mathcal{O}(n+1)$	0.842 ± 0.006	0.860 ± 0.006	0.844 ± 0.009	0.750 ± 0.008	0.796 ± 0.010	0.771 ± 0.032	0.780 ± 0.013	0.806 ± 0.009
	$\mathcal{O}(2^n)$	0.838 ± 0.005	0.853 ± 0.004	0.843 ± 0.006	0.770 ± 0.005	0.815 ± 0.010	0.818 ± 0.009	0.777 ± 0.006	0.816 ± 0.002
Qwen2.5-Omni-7B									
Linear	$\mathcal{O}(1)$	0.853 ± 0.009	0.859 ± 0.009	0.749 ± 0.023	0.712 ± 0.014	0.759 ± 0.004	0.697 ± 0.012	0.731 ± 0.011	0.766 ± 0.009
	$\mathcal{O}(1)$ **- Ours**	0.819 ± 0.002	0.844 ± 0.003	0.741 ± 0.014	0.668 ± 0.003	0.781 ± 0.004	0.697 ± 0.009	0.691 ± 0.003	0.749 ± 0.003
	$\mathcal{O}(n)$	0.858 ± 0.003	0.866 ± 0.004	0.778 ± 0.016	0.477 ± 0.015	0.786 ± 0.007	0.740 ± 0.013	0.510 ± 0.009	0.716 ± 0.005
	$\mathcal{O}(n+1)$	0.846 ± 0.010	0.865 ± 0.002	0.787 ± 0.014	0.759 ± 0.012	0.826 ± 0.002	0.831 ± 0.005	0.775 ± 0.010	0.813 ± 0.003
	$\mathcal{O}(2^n)$	0.853 ± 0.009	0.866 ± 0.003	0.725 ± 0.011	0.780 ± 0.006	0.822 ± 0.002	0.830 ± 0.003	0.783 ± 0.005	0.808 ± 0.003
MLP	$\mathcal{O}(1)$	0.857 ± 0.009	0.845 ± 0.019	0.781 ± 0.024	0.554 ± 0.064	0.794 ± 0.013	0.684 ± 0.014	0.598 ± 0.060	0.730 ± 0.018
	$\mathcal{O}(1)$ **- Ours**	0.848 ± 0.006	0.855 ± 0.004	0.799 ± 0.008	0.707 ± 0.008	0.809 ± 0.006	0.745 ± 0.007	0.718 ± 0.004	0.783 ± 0.004
	$\mathcal{O}(n)$	0.859 ± 0.004	0.873 ± 0.004	0.796 ± 0.008	0.558 ± 0.045	0.806 ± 0.006	0.772 ± 0.003	0.606 ± 0.040	0.753 ± 0.014
	$\mathcal{O}(n+1)$	0.856 ± 0.012	0.872 ± 0.004	0.789 ± 0.011	0.784 ± 0.010	0.830 ± 0.005	0.835 ± 0.009	0.799 ± 0.010	0.824 ± 0.004
	$\mathcal{O}(2^n)$	0.857 ± 0.009	0.865 ± 0.002	0.723 ± 0.026	0.784 ± 0.007	0.824 ± 0.007	0.835 ± 0.005	0.791 ± 0.006	0.811 ± 0.003
Qwen3-Omni-30B									
Linear	$\mathcal{O}(1)$	0.848 ± 0.004	0.559 ± 0.133	0.395 ± 0.022	0.358 ± 0.257	0.816 ± 0.010	0.275 ± 0.088	0.378 ± 0.226	0.518 ± 0.100
	$\mathcal{O}(1)$ **- Ours**	0.836 ± 0.002	0.731 ± 0.049	0.524 ± 0.030	0.468 ± 0.077	0.829 ± 0.003	0.360 ± 0.121	0.491 ± 0.086	0.606 ± 0.046
	$\mathcal{O}(n)$	0.847 ± 0.005	0.830 ± 0.013	0.852 ± 0.003	0.544 ± 0.056	0.829 ± 0.004	0.760 ± 0.024	0.584 ± 0.057	0.750 ± 0.021
	$\mathcal{O}(n+1)$	0.841 ± 0.002	0.845 ± 0.007	0.854 ± 0.005	0.782 ± 0.007	0.841 ± 0.003	0.802 ± 0.006	0.786 ± 0.003	0.822 ± 0.002
	$\mathcal{O}(2^n)$	0.848 ± 0.004	0.821 ± 0.004	0.845 ± 0.002	0.814 ± 0.003	0.833 ± 0.003	0.820 ± 0.006	0.796 ± 0.003	0.825 ± 0.001
MLP	$\mathcal{O}(1)$	0.852 ± 0.008	0.390 ± 0.162	0.388 ± 0.112	0.342 ± 0.177	0.817 ± 0.014	0.221 ± 0.212	0.314 ± 0.181	0.475 ± 0.118
	$\mathcal{O}(1)$ **- Ours**	0.851 ± 0.004	0.753 ± 0.062	0.629 ± 0.034	0.573 ± 0.114	0.840 ± 0.003	0.563 ± 0.218	0.571 ± 0.121	0.683 ± 0.076
	$\mathcal{O}(n)$	0.854 ± 0.005	0.834 ± 0.014	0.856 ± 0.008	0.632 ± 0.065	0.833 ± 0.007	0.632 ± 0.056	0.665 ± 0.043	0.758 ± 0.020
	$\mathcal{O}(n+1)$	0.857 ± 0.007	0.851 ± 0.011	0.860 ± 0.007	0.784 ± 0.013	0.842 ± 0.004	0.793 ± 0.017	0.799 ± 0.006	0.826 ± 0.005
	$\mathcal{O}(2^n)$	0.852 ± 0.008	0.838 ± 0.009	0.851 ± 0.006	0.814 ± 0.010	0.837 ± 0.009	0.819 ± 0.003	0.798 ± 0.008	0.830 ± 0.005
Phi-4-multimodal									
Linear	$\mathcal{O}(1)$	0.849 ± 0.004	0.652 ± 0.074	0.250 ± 0.026	0.093 ± 0.044	0.688 ± 0.040	0.067 ± 0.031	0.079 ± 0.038	0.383 ± 0.024
	$\mathcal{O}(1)$ **- Ours**	0.817 ± 0.011	0.521 ± 0.047	0.169 ± 0.019	0.172 ± 0.018	0.503 ± 0.021	0.105 ± 0.017	0.163 ± 0.019	0.350 ± 0.020
	$\mathcal{O}(n)$	0.854 ± 0.008	0.836 ± 0.005	0.850 ± 0.005	0.539 ± 0.041	0.708 ± 0.020	0.545 ± 0.040	0.556 ± 0.038	0.698 ± 0.019
	$\mathcal{O}(n+1)$	0.856 ± 0.004	0.818 ± 0.005	0.852 ± 0.002	0.726 ± 0.009	0.731 ± 0.014	0.755 ± 0.003	0.743 ± 0.005	0.783 ± 0.002
	$\mathcal{O}(2^n)$	0.849 ± 0.004	0.817 ± 0.007	0.846 ± 0.003	0.761 ± 0.004	0.794 ± 0.006	0.806 ± 0.004	0.761 ± 0.005	0.805 ± 0.002
MLP	$\mathcal{O}(1)$	0.855 ± 0.005	0.588 ± 0.126	0.314 ± 0.133	0.048 ± 0.051	0.670 ± 0.030	0.048 ± 0.048	0.053 ± 0.056	0.368 ± 0.048
	$\mathcal{O}(1)$ **- Ours**	0.852 ± 0.005	0.550 ± 0.021	0.226 ± 0.023	0.272 ± 0.053	0.592 ± 0.023	0.201 ± 0.041	0.252 ± 0.054	0.421 ± 0.029
	$\mathcal{O}(n)$	0.859 ± 0.004	0.827 ± 0.012	0.854 ± 0.004	0.447 ± 0.105	0.736 ± 0.008	0.469 ± 0.115	0.470 ± 0.107	0.666 ± 0.050
	$\mathcal{O}(n+1)$	0.867 ± 0.005	0.828 ± 0.006	0.848 ± 0.006	0.705 ± 0.033	0.738 ± 0.013	0.766 ± 0.006	0.752 ± 0.011	0.786 ± 0.005
	$\mathcal{O}(2^n)$	0.855 ± 0.005	0.827 ± 0.008	0.847 ± 0.007	0.754 ± 0.006	0.791 ± 0.012	0.809 ± 0.003	0.760 ± 0.005	0.806 ± 0.003
Interactive-Omni-8B									
Linear	$\mathcal{O}(1)$	0.851 ± 0.003	0.716 ± 0.041	0.094 ± 0.056	0.254 ± 0.072	0.772 ± 0.015	0.104 ± 0.052	0.403 ± 0.098	0.456 ± 0.046
	$\mathcal{O}(1)$ **- Ours**	0.824 ± 0.010	0.836 ± 0.006	0.591 ± 0.027	0.757 ± 0.003	0.776 ± 0.011	0.727 ± 0.019	0.755 ± 0.003	0.752 ± 0.008
	$\mathcal{O}(n)$	0.848 ± 0.005	0.850 ± 0.008	0.851 ± 0.005	0.694 ± 0.014	0.808 ± 0.003	0.824 ± 0.004	0.778 ± 0.004	0.808 ± 0.003
	$\mathcal{O}(n+1)$	0.840 ± 0.006	0.841 ± 0.007	0.847 ± 0.004	0.758 ± 0.008	0.819 ± 0.005	0.826 ± 0.002	0.810 ± 0.008	0.820 ± 0.002
	$\mathcal{O}(2^n)$	0.851 ± 0.003	0.839 ± 0.005	0.846 ± 0.007	0.813 ± 0.002	0.814 ± 0.010	0.827 ± 0.008	0.804 ± 0.004	0.828 ± 0.002
MLP	$\mathcal{O}(1)$	0.853 ± 0.002	0.758 ± 0.043	0.257 ± 0.140	0.477 ± 0.103	0.787 ± 0.022	0.259 ± 0.119	0.573 ± 0.086	0.566 ± 0.065
	$\mathcal{O}(1)$ **- Ours**	0.849 ± 0.003	0.841 ± 0.005	0.693 ± 0.099	0.756 ± 0.006	0.802 ± 0.004	0.740 ± 0.031	0.766 ± 0.011	0.778 ± 0.020
	$\mathcal{O}(n)$	0.851 ± 0.009	0.855 ± 0.005	0.850 ± 0.007	0.678 ± 0.043	0.806 ± 0.006	0.813 ± 0.015	0.777 ± 0.004	0.804 ± 0.009
	$\mathcal{O}(n+1)$	0.860 ± 0.006	0.852 ± 0.005	0.847 ± 0.005	0.762 ± 0.022	0.829 ± 0.003	0.834 ± 0.005	0.818 ± 0.007	0.829 ± 0.004
	$\mathcal{O}(2^n)$	0.853 ± 0.002	0.838 ± 0.004	0.849 ± 0.009	0.807 ± 0.006	0.817 ± 0.016	0.830 ± 0.010	0.808 ± 0.011	0.829 ± 0.007

Table 32: **Mean AUC for the three supervision regimes on Aegis-Omni.** Entries report mean ± sample standard deviation across five seeds, rounded to three decimals.

Monitor	Method ↓ / Test →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	$\mathcal{O}(1)$	0.901 ± 0.001	0.908 ± 0.002	0.522 ± 0.014	0.745 ± 0.007	0.728 ± 0.017	0.687 ± 0.013	0.738 ± 0.009	0.747 ± 0.007
	$\mathcal{O}(1)$ **- Ours**	0.885 ± 0.005	0.882 ± 0.005	0.713 ± 0.010	0.795 ± 0.005	0.824 ± 0.006	0.780 ± 0.003	0.795 ± 0.005	0.810 ± 0.002
	$\mathcal{O}(n)$	0.901 ± 0.001	0.908 ± 0.002	0.886 ± 0.001	0.697 ± 0.009	0.834 ± 0.002	0.808 ± 0.006	0.719 ± 0.009	0.822 ± 0.004
	$\mathcal{O}(n+1)$	0.897 ± 0.001	0.903 ± 0.001	0.888 ± 0.001	0.817 ± 0.003	0.848 ± 0.001	0.833 ± 0.006	0.827 ± 0.002	0.859 ± 0.001
	$\mathcal{O}(2^n)$	0.901 ± 0.001	0.915 ± 0.001	0.892 ± 0.001	0.829 ± 0.003	0.868 ± 0.002	0.880 ± 0.002	0.841 ± 0.003	0.875 ± 0.001
MLP	$\mathcal{O}(1)$	0.903 ± 0.003	0.905 ± 0.004	0.536 ± 0.053	0.752 ± 0.015	0.721 ± 0.024	0.729 ± 0.013	0.746 ± 0.015	0.756 ± 0.009
	$\mathcal{O}(1)$ **- Ours**	0.901 ± 0.004	0.891 ± 0.005	0.675 ± 0.025	0.794 ± 0.008	0.827 ± 0.008	0.781 ± 0.006	0.794 ± 0.007	0.809 ± 0.006
	$\mathcal{O}(n)$	0.905 ± 0.002	0.914 ± 0.006	0.888 ± 0.007	0.704 ± 0.017	0.832 ± 0.005	0.816 ± 0.012	0.723 ± 0.017	0.826 ± 0.008
	$\mathcal{O}(n+1)$	0.902 ± 0.003	0.911 ± 0.004	0.888 ± 0.005	0.824 ± 0.007	0.843 ± 0.007	0.838 ± 0.014	0.839 ± 0.007	0.863 ± 0.006
	$\mathcal{O}(2^n)$	0.903 ± 0.003	0.916 ± 0.002	0.894 ± 0.002	0.834 ± 0.002	0.872 ± 0.003	0.887 ± 0.001	0.847 ± 0.002	0.879 ± 0.001
Qwen2.5-Omni-7B									
Linear	$\mathcal{O}(1)$	0.905 ± 0.002	0.908 ± 0.002	0.672 ± 0.022	0.751 ± 0.014	0.777 ± 0.012	0.772 ± 0.012	0.760 ± 0.012	0.792 ± 0.006
	$\mathcal{O}(1)$ **- Ours**	0.895 ± 0.005	0.890 ± 0.006	0.767 ± 0.009	0.790 ± 0.007	0.847 ± 0.002	0.791 ± 0.009	0.791 ± 0.007	0.825 ± 0.001
	$\mathcal{O}(n)$	0.901 ± 0.002	0.909 ± 0.002	0.879 ± 0.005	0.756 ± 0.008	0.838 ± 0.005	0.843 ± 0.004	0.771 ± 0.008	0.842 ± 0.004
	$\mathcal{O}(n+1)$	0.901 ± 0.001	0.911 ± 0.002	0.879 ± 0.005	0.833 ± 0.001	0.851 ± 0.001	0.846 ± 0.008	0.844 ± 0.001	0.866 ± 0.002
	$\mathcal{O}(2^n)$	0.905 ± 0.002	0.913 ± 0.003	0.890 ± 0.002	0.837 ± 0.002	0.873 ± 0.004	0.882 ± 0.003	0.849 ± 0.002	0.878 ± 0.001
MLP	$\mathcal{O}(1)$	0.906 ± 0.002	0.901 ± 0.004	0.687 ± 0.044	0.736 ± 0.011	0.791 ± 0.023	0.781 ± 0.013	0.748 ± 0.012	0.793 ± 0.013
	$\mathcal{O}(1)$ **- Ours**	0.901 ± 0.003	0.890 ± 0.005	0.735 ± 0.010	0.769 ± 0.009	0.843 ± 0.002	0.796 ± 0.004	0.775 ± 0.007	0.816 ± 0.003
	$\mathcal{O}(n)$	0.899 ± 0.001	0.904 ± 0.002	0.878 ± 0.005	0.729 ± 0.007	0.826 ± 0.007	0.832 ± 0.004	0.747 ± 0.006	0.831 ± 0.003
	$\mathcal{O}(n+1)$	0.900 ± 0.003	0.910 ± 0.002	0.876 ± 0.008	0.831 ± 0.007	0.848 ± 0.009	0.849 ± 0.005	0.842 ± 0.005	0.865 ± 0.003
	$\mathcal{O}(2^n)$	0.906 ± 0.002	0.914 ± 0.001	0.891 ± 0.002	0.839 ± 0.002	0.877 ± 0.003	0.887 ± 0.003	0.850 ± 0.003	0.881 ± 0.001
Qwen3-Omni-30B									
Linear	$\mathcal{O}(1)$	0.909 ± 0.002	0.792 ± 0.016	0.726 ± 0.012	0.780 ± 0.024	0.849 ± 0.013	0.681 ± 0.023	0.757 ± 0.024	0.785 ± 0.012
	$\mathcal{O}(1)$ **- Ours**	0.894 ± 0.002	0.876 ± 0.001	0.756 ± 0.003	0.833 ± 0.002	0.855 ± 0.001	0.785 ± 0.005	0.817 ± 0.002	0.831 ± 0.001
	$\mathcal{O}(n)$	0.901 ± 0.001	0.900 ± 0.001	0.880 ± 0.001	0.824 ± 0.002	0.864 ± 0.001	0.828 ± 0.004	0.821 ± 0.003	0.860 ± 0.001
	$\mathcal{O}(n+1)$	0.901 ± 0.001	0.902 ± 0.001	0.878 ± 0.001	0.846 ± 0.002	0.869 ± 0.002	0.832 ± 0.002	0.848 ± 0.002	0.868 ± 0.001
	$\mathcal{O}(2^n)$	0.909 ± 0.002	0.905 ± 0.002	0.890 ± 0.002	0.869 ± 0.001	0.884 ± 0.001	0.871 ± 0.001	0.861 ± 0.001	0.884 ± 0.001
MLP	$\mathcal{O}(1)$	0.908 ± 0.003	0.801 ± 0.029	0.707 ± 0.045	0.799 ± 0.009	0.821 ± 0.023	0.697 ± 0.023	0.761 ± 0.017	0.785 ± 0.019
	$\mathcal{O}(1)$ **- Ours**	0.906 ± 0.002	0.875 ± 0.004	0.731 ± 0.010	0.805 ± 0.015	0.848 ± 0.002	0.743 ± 0.011	0.785 ± 0.014	0.813 ± 0.006
	$\mathcal{O}(n)$	0.905 ± 0.002	0.905 ± 0.002	0.892 ± 0.001	0.821 ± 0.002	0.864 ± 0.002	0.810 ± 0.007	0.819 ± 0.003	0.860 ± 0.001
	$\mathcal{O}(n+1)$	0.903 ± 0.003	0.907 ± 0.002	0.889 ± 0.003	0.859 ± 0.003	0.864 ± 0.002	0.833 ± 0.006	0.858 ± 0.002	0.873 ± 0.002
	$\mathcal{O}(2^n)$	0.908 ± 0.003	0.906 ± 0.002	0.893 ± 0.003	0.873 ± 0.004	0.887 ± 0.002	0.873 ± 0.003	0.861 ± 0.005	0.886 ± 0.002
Phi-4-multimodal									
Linear	$\mathcal{O}(1)$	0.904 ± 0.002	0.850 ± 0.006	0.626 ± 0.039	0.688 ± 0.005	0.760 ± 0.014	0.678 ± 0.008	0.693 ± 0.005	0.743 ± 0.005
	$\mathcal{O}(1)$ **- Ours**	0.899 ± 0.002	0.864 ± 0.002	0.818 ± 0.003	0.753 ± 0.005	0.828 ± 0.002	0.760 ± 0.004	0.759 ± 0.004	0.812 ± 0.003
	$\mathcal{O}(n)$	0.903 ± 0.001	0.887 ± 0.002	0.890 ± 0.002	0.712 ± 0.005	0.787 ± 0.011	0.734 ± 0.003	0.723 ± 0.004	0.805 ± 0.003
	$\mathcal{O}(n+1)$	0.898 ± 0.002	0.888 ± 0.001	0.890 ± 0.002	0.807 ± 0.001	0.796 ± 0.006	0.826 ± 0.002	0.817 ± 0.002	0.846 ± 0.002
	$\mathcal{O}(2^n)$	0.904 ± 0.002	0.893 ± 0.001	0.892 ± 0.002	0.810 ± 0.005	0.851 ± 0.003	0.853 ± 0.005	0.822 ± 0.004	0.861 ± 0.001
MLP	$\mathcal{O}(1)$	0.903 ± 0.004	0.839 ± 0.011	0.659 ± 0.044	0.686 ± 0.011	0.756 ± 0.022	0.690 ± 0.012	0.694 ± 0.011	0.747 ± 0.008
	$\mathcal{O}(1)$ **- Ours**	0.897 ± 0.001	0.844 ± 0.002	0.746 ± 0.016	0.708 ± 0.012	0.782 ± 0.003	0.713 ± 0.010	0.711 ± 0.011	0.771 ± 0.006
	$\mathcal{O}(n)$	0.902 ± 0.006	0.895 ± 0.002	0.892 ± 0.002	0.728 ± 0.013	0.786 ± 0.016	0.738 ± 0.009	0.735 ± 0.012	0.811 ± 0.008
	$\mathcal{O}(n+1)$	0.900 ± 0.002	0.896 ± 0.003	0.891 ± 0.002	0.814 ± 0.001	0.784 ± 0.005	0.835 ± 0.001	0.825 ± 0.001	0.849 ± 0.001
	$\mathcal{O}(2^n)$	0.903 ± 0.004	0.897 ± 0.001	0.893 ± 0.003	0.815 ± 0.005	0.851 ± 0.003	0.854 ± 0.007	0.824 ± 0.006	0.862 ± 0.002
Interactive-Omni-8B									
Linear	$\mathcal{O}(1)$	0.910 ± 0.001	0.907 ± 0.002	0.728 ± 0.015	0.790 ± 0.015	0.800 ± 0.004	0.806 ± 0.006	0.804 ± 0.007	0.821 ± 0.005
	$\mathcal{O}(1)$ **- Ours**	0.900 ± 0.002	0.889 ± 0.001	0.791 ± 0.002	0.853 ± 0.002	0.859 ± 0.003	0.831 ± 0.002	0.843 ± 0.001	0.852 ± 0.001
	$\mathcal{O}(n)$	0.907 ± 0.002	0.905 ± 0.003	0.885 ± 0.003	0.770 ± 0.013	0.846 ± 0.005	0.843 ± 0.007	0.799 ± 0.009	0.851 ± 0.005
	$\mathcal{O}(n+1)$	0.907 ± 0.002	0.904 ± 0.007	0.882 ± 0.006	0.830 ± 0.008	0.861 ± 0.002	0.859 ± 0.006	0.848 ± 0.006	0.870 ± 0.005
	$\mathcal{O}(2^n)$	0.910 ± 0.001	0.915 ± 0.002	0.892 ± 0.002	0.860 ± 0.002	0.874 ± 0.004	0.875 ± 0.001	0.854 ± 0.003	0.883 ± 0.001
MLP	$\mathcal{O}(1)$	0.911 ± 0.004	0.905 ± 0.002	0.739 ± 0.023	0.776 ± 0.020	0.809 ± 0.008	0.821 ± 0.017	0.806 ± 0.009	0.824 ± 0.007
	$\mathcal{O}(1)$ **- Ours**	0.909 ± 0.002	0.887 ± 0.002	0.740 ± 0.005	0.846 ± 0.003	0.853 ± 0.002	0.827 ± 0.003	0.837 ± 0.002	0.843 ± 0.001
	$\mathcal{O}(n)$	0.911 ± 0.001	0.907 ± 0.005	0.889 ± 0.007	0.772 ± 0.015	0.844 ± 0.006	0.851 ± 0.009	0.806 ± 0.014	0.854 ± 0.007
	$\mathcal{O}(n+1)$	0.909 ± 0.006	0.902 ± 0.005	0.885 ± 0.010	0.830 ± 0.007	0.859 ± 0.007	0.858 ± 0.015	0.853 ± 0.007	0.871 ± 0.008
	$\mathcal{O}(2^n)$	0.911 ± 0.004	0.913 ± 0.005	0.893 ± 0.008	0.864 ± 0.002	0.882 ± 0.003	0.881 ± 0.001	0.860 ± 0.001	0.886 ± 0.001

Table 33: **Mean accuracy for the three supervision regimes on Aegis-Omni.** Entries report mean ± sample standard deviation across five seeds, rounded to three decimals.

Monitor	Method ↓ / Test →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	$\mathcal{O}(1)$	0.826 ± 0.003	0.830 ± 0.007	0.468 ± 0.002	0.631 ± 0.040	0.597 ± 0.025	0.561 ± 0.032	0.614 ± 0.039	0.647 ± 0.019
	$\mathcal{O}(1)$ **- Ours**	0.811 ± 0.003	0.795 ± 0.011	0.550 ± 0.007	0.680 ± 0.016	0.665 ± 0.010	0.655 ± 0.015	0.672 ± 0.015	0.690 ± 0.005
	$\mathcal{O}(n)$	0.826 ± 0.003	0.834 ± 0.003	0.822 ± 0.004	0.588 ± 0.010	0.713 ± 0.009	0.658 ± 0.016	0.590 ± 0.011	0.719 ± 0.005
	$\mathcal{O}(n+1)$	0.820 ± 0.002	0.830 ± 0.003	0.822 ± 0.002	0.747 ± 0.002	0.748 ± 0.006	0.738 ± 0.023	0.754 ± 0.003	0.780 ± 0.003
	$\mathcal{O}(2^n)$	0.826 ± 0.003	0.837 ± 0.003	0.828 ± 0.004	0.754 ± 0.005	0.794 ± 0.001	0.802 ± 0.003	0.764 ± 0.002	0.801 ± 0.001
MLP	$\mathcal{O}(1)$	0.829 ± 0.008	0.810 ± 0.028	0.466 ± 0.002	0.613 ± 0.058	0.579 ± 0.026	0.545 ± 0.049	0.597 ± 0.059	0.634 ± 0.031
	$\mathcal{O}(1)$ **- Ours**	0.824 ± 0.005	0.812 ± 0.008	0.521 ± 0.027	0.683 ± 0.015	0.670 ± 0.014	0.658 ± 0.010	0.676 ± 0.011	0.692 ± 0.005
	$\mathcal{O}(n)$	0.828 ± 0.005	0.837 ± 0.007	0.818 ± 0.002	0.564 ± 0.033	0.701 ± 0.011	0.661 ± 0.042	0.572 ± 0.032	0.712 ± 0.013
	$\mathcal{O}(n+1)$	0.824 ± 0.004	0.829 ± 0.004	0.819 ± 0.005	0.750 ± 0.008	0.730 ± 0.025	0.706 ± 0.027	0.763 ± 0.006	0.775 ± 0.008
	$\mathcal{O}(2^n)$	0.829 ± 0.008	0.838 ± 0.003	0.829 ± 0.003	0.757 ± 0.002	0.798 ± 0.004	0.810 ± 0.002	0.767 ± 0.004	0.804 ± 0.002
Qwen2.5-Omni-7B									
Linear	$\mathcal{O}(1)$	0.831 ± 0.004	0.827 ± 0.009	0.511 ± 0.038	0.534 ± 0.009	0.667 ± 0.013	0.600 ± 0.012	0.541 ± 0.007	0.644 ± 0.009
	$\mathcal{O}(1)$ **- Ours**	0.822 ± 0.004	0.819 ± 0.004	0.593 ± 0.012	0.612 ± 0.020	0.720 ± 0.005	0.646 ± 0.010	0.616 ± 0.022	0.690 ± 0.009
	$\mathcal{O}(n)$	0.826 ± 0.003	0.835 ± 0.004	0.821 ± 0.007	0.574 ± 0.007	0.731 ± 0.007	0.732 ± 0.008	0.585 ± 0.006	0.729 ± 0.005
	$\mathcal{O}(n+1)$	0.827 ± 0.004	0.831 ± 0.003	0.815 ± 0.006	0.751 ± 0.004	0.770 ± 0.004	0.767 ± 0.012	0.762 ± 0.003	0.789 ± 0.002
	$\mathcal{O}(2^n)$	0.831 ± 0.004	0.842 ± 0.002	0.825 ± 0.002	0.761 ± 0.005	0.802 ± 0.001	0.807 ± 0.006	0.771 ± 0.002	0.806 ± 0.001
MLP	$\mathcal{O}(1)$	0.831 ± 0.006	0.819 ± 0.014	0.487 ± 0.018	0.535 ± 0.016	0.621 ± 0.023	0.569 ± 0.021	0.538 ± 0.015	0.629 ± 0.012
	$\mathcal{O}(1)$ **- Ours**	0.833 ± 0.004	0.824 ± 0.007	0.548 ± 0.013	0.605 ± 0.022	0.696 ± 0.008	0.647 ± 0.011	0.611 ± 0.020	0.681 ± 0.008
	$\mathcal{O}(n)$	0.829 ± 0.005	0.833 ± 0.008	0.816 ± 0.003	0.562 ± 0.016	0.719 ± 0.022	0.707 ± 0.020	0.573 ± 0.017	0.720 ± 0.008
	$\mathcal{O}(n+1)$	0.828 ± 0.005	0.828 ± 0.004	0.812 ± 0.008	0.746 ± 0.014	0.767 ± 0.009	0.770 ± 0.004	0.761 ± 0.009	0.787 ± 0.004
	$\mathcal{O}(2^n)$	0.831 ± 0.006	0.843 ± 0.006	0.825 ± 0.005	0.757 ± 0.005	0.798 ± 0.003	0.809 ± 0.002	0.766 ± 0.006	0.804 ± 0.003
Qwen3-Omni-30B									
Linear	$\mathcal{O}(1)$	0.835 ± 0.002	0.640 ± 0.064	0.672 ± 0.015	0.606 ± 0.052	0.768 ± 0.011	0.569 ± 0.043	0.603 ± 0.052	0.670 ± 0.028
	$\mathcal{O}(1)$ **- Ours**	0.823 ± 0.001	0.794 ± 0.011	0.537 ± 0.007	0.576 ± 0.021	0.730 ± 0.003	0.546 ± 0.014	0.561 ± 0.018	0.653 ± 0.010
	$\mathcal{O}(n)$	0.830 ± 0.002	0.830 ± 0.004	0.826 ± 0.004	0.619 ± 0.027	0.761 ± 0.007	0.700 ± 0.033	0.610 ± 0.018	0.740 ± 0.010
	$\mathcal{O}(n+1)$	0.828 ± 0.002	0.829 ± 0.003	0.823 ± 0.003	0.769 ± 0.007	0.784 ± 0.005	0.762 ± 0.005	0.767 ± 0.005	0.795 ± 0.004
	$\mathcal{O}(2^n)$	0.835 ± 0.002	0.837 ± 0.004	0.828 ± 0.004	0.797 ± 0.003	0.809 ± 0.002	0.796 ± 0.003	0.789 ± 0.004	0.813 ± 0.002
MLP	$\mathcal{O}(1)$	0.833 ± 0.004	0.619 ± 0.114	0.572 ± 0.051	0.619 ± 0.094	0.723 ± 0.008	0.572 ± 0.076	0.599 ± 0.088	0.648 ± 0.057
	$\mathcal{O}(1)$ **- Ours**	0.830 ± 0.002	0.730 ± 0.065	0.496 ± 0.008	0.565 ± 0.044	0.709 ± 0.011	0.516 ± 0.017	0.548 ± 0.033	0.628 ± 0.022
	$\mathcal{O}(n)$	0.829 ± 0.004	0.834 ± 0.004	0.829 ± 0.003	0.601 ± 0.023	0.755 ± 0.008	0.644 ± 0.021	0.604 ± 0.014	0.728 ± 0.006
	$\mathcal{O}(n+1)$	0.829 ± 0.003	0.830 ± 0.006	0.826 ± 0.005	0.776 ± 0.008	0.767 ± 0.014	0.759 ± 0.004	0.776 ± 0.005	0.795 ± 0.003
	$\mathcal{O}(2^n)$	0.833 ± 0.004	0.836 ± 0.005	0.828 ± 0.004	0.794 ± 0.005	0.807 ± 0.008	0.794 ± 0.007	0.785 ± 0.003	0.811 ± 0.002
Phi-4-multimodal									
Linear	$\mathcal{O}(1)$	0.830 ± 0.005	0.709 ± 0.021	0.507 ± 0.048	0.476 ± 0.003	0.642 ± 0.027	0.472 ± 0.002	0.475 ± 0.004	0.587 ± 0.010
	$\mathcal{O}(1)$ **- Ours**	0.826 ± 0.003	0.741 ± 0.006	0.532 ± 0.006	0.515 ± 0.003	0.530 ± 0.005	0.513 ± 0.005	0.513 ± 0.003	0.596 ± 0.003
	$\mathcal{O}(n)$	0.829 ± 0.002	0.822 ± 0.004	0.827 ± 0.003	0.553 ± 0.012	0.654 ± 0.009	0.560 ± 0.016	0.558 ± 0.015	0.686 ± 0.007
	$\mathcal{O}(n+1)$	0.821 ± 0.003	0.815 ± 0.004	0.827 ± 0.003	0.737 ± 0.004	0.681 ± 0.006	0.755 ± 0.005	0.742 ± 0.002	0.768 ± 0.002
	$\mathcal{O}(2^n)$	0.830 ± 0.005	0.830 ± 0.003	0.827 ± 0.004	0.736 ± 0.006	0.779 ± 0.003	0.782 ± 0.008	0.743 ± 0.007	0.790 ± 0.002
MLP	$\mathcal{O}(1)$	0.828 ± 0.004	0.691 ± 0.065	0.536 ± 0.065	0.470 ± 0.010	0.629 ± 0.039	0.468 ± 0.008	0.469 ± 0.010	0.584 ± 0.013
	$\mathcal{O}(1)$ **- Ours**	0.828 ± 0.004	0.681 ± 0.006	0.495 ± 0.012	0.503 ± 0.011	0.517 ± 0.009	0.497 ± 0.009	0.501 ± 0.010	0.575 ± 0.004
	$\mathcal{O}(n)$	0.827 ± 0.004	0.826 ± 0.005	0.828 ± 0.003	0.535 ± 0.024	0.643 ± 0.016	0.533 ± 0.019	0.537 ± 0.025	0.675 ± 0.012
	$\mathcal{O}(n+1)$	0.828 ± 0.004	0.821 ± 0.005	0.827 ± 0.007	0.727 ± 0.014	0.657 ± 0.014	0.760 ± 0.005	0.746 ± 0.008	0.767 ± 0.005
	$\mathcal{O}(2^n)$	0.828 ± 0.004	0.830 ± 0.007	0.825 ± 0.002	0.739 ± 0.006	0.779 ± 0.006	0.781 ± 0.010	0.746 ± 0.005	0.790 ± 0.001
Interactive-Omni-8B									
Linear	$\mathcal{O}(1)$	0.834 ± 0.003	0.802 ± 0.018	0.552 ± 0.063	0.593 ± 0.028	0.637 ± 0.035	0.545 ± 0.034	0.576 ± 0.033	0.648 ± 0.028
	$\mathcal{O}(1)$ **- Ours**	0.821 ± 0.001	0.812 ± 0.003	0.620 ± 0.008	0.715 ± 0.007	0.705 ± 0.005	0.706 ± 0.004	0.698 ± 0.003	0.725 ± 0.003
	$\mathcal{O}(n)$	0.831 ± 0.004	0.827 ± 0.005	0.819 ± 0.003	0.610 ± 0.014	0.730 ± 0.014	0.729 ± 0.021	0.671 ± 0.025	0.745 ± 0.011
	$\mathcal{O}(n+1)$	0.829 ± 0.002	0.827 ± 0.006	0.815 ± 0.005	0.717 ± 0.017	0.777 ± 0.006	0.781 ± 0.007	0.766 ± 0.006	0.787 ± 0.003
	$\mathcal{O}(2^n)$	0.834 ± 0.003	0.835 ± 0.004	0.830 ± 0.004	0.782 ± 0.004	0.799 ± 0.004	0.805 ± 0.005	0.774 ± 0.003	0.808 ± 0.000
MLP	$\mathcal{O}(1)$	0.829 ± 0.008	0.768 ± 0.036	0.501 ± 0.016	0.563 ± 0.028	0.603 ± 0.025	0.521 ± 0.029	0.552 ± 0.034	0.620 ± 0.023
	$\mathcal{O}(1)$ **- Ours**	0.827 ± 0.002	0.805 ± 0.003	0.573 ± 0.012	0.716 ± 0.021	0.702 ± 0.009	0.693 ± 0.012	0.712 ± 0.008	0.718 ± 0.008
	$\mathcal{O}(n)$	0.830 ± 0.003	0.834 ± 0.004	0.819 ± 0.008	0.609 ± 0.016	0.744 ± 0.010	0.740 ± 0.040	0.678 ± 0.025	0.751 ± 0.013
	$\mathcal{O}(n+1)$	0.832 ± 0.003	0.823 ± 0.006	0.817 ± 0.007	0.720 ± 0.021	0.766 ± 0.011	0.771 ± 0.010	0.770 ± 0.006	0.786 ± 0.006
	$\mathcal{O}(2^n)$	0.829 ± 0.008	0.836 ± 0.006	0.826 ± 0.010	0.780 ± 0.006	0.803 ± 0.001	0.805 ± 0.004	0.779 ± 0.003	0.808 ± 0.002

Table 34: **Mean f1 score for the three supervision regimes on Aegis-Omni.** Entries report mean ± sample standard deviation across five seeds, rounded to three decimals.

Monitor	Method ↓ / Test →	Text	Image	Audio	T+I	T+A	A+I	T+A+I	Avg.
Qwen2.5-Omni-3B									
Linear	$\mathcal{O}(1)$	0.836 ± 0.003	0.843 ± 0.015	0.070 ± 0.033	0.535 ± 0.115	0.435 ± 0.059	0.349 ± 0.102	0.487 ± 0.115	0.508 ± 0.056
	$\mathcal{O}(1)$ **- Ours**	0.821 ± 0.002	0.821 ± 0.004	0.382 ± 0.020	0.649 ± 0.041	0.590 ± 0.016	0.586 ± 0.034	0.629 ± 0.039	0.640 ± 0.017
	$\mathcal{O}(n)$	0.836 ± 0.003	0.846 ± 0.004	0.840 ± 0.004	0.419 ± 0.033	0.682 ± 0.015	0.584 ± 0.036	0.423 ± 0.037	0.661 ± 0.013
	$\mathcal{O}(n+1)$	0.833 ± 0.005	0.844 ± 0.003	0.839 ± 0.002	0.766 ± 0.006	0.737 ± 0.010	0.734 ± 0.035	0.774 ± 0.006	0.790 ± 0.006
	$\mathcal{O}(2^n)$	0.836 ± 0.003	0.848 ± 0.003	0.843 ± 0.004	0.777 ± 0.006	0.813 ± 0.002	0.818 ± 0.003	0.787 ± 0.002	0.817 ± 0.001
MLP	$\mathcal{O}(1)$	0.839 ± 0.009	0.812 ± 0.047	0.043 ± 0.026	0.478 ± 0.159	0.385 ± 0.076	0.286 ± 0.154	0.434 ± 0.166	0.468 ± 0.087
	$\mathcal{O}(1)$ **- Ours**	0.836 ± 0.004	0.834 ± 0.003	0.308 ± 0.076	0.652 ± 0.046	0.599 ± 0.022	0.604 ± 0.027	0.636 ± 0.037	0.638 ± 0.012
	$\mathcal{O}(n)$	0.841 ± 0.005	0.848 ± 0.008	0.833 ± 0.006	0.346 ± 0.097	0.660 ± 0.024	0.584 ± 0.093	0.368 ± 0.094	0.640 ± 0.036
	$\mathcal{O}(n+1)$	0.837 ± 0.005	0.843 ± 0.005	0.837 ± 0.006	0.767 ± 0.013	0.711 ± 0.043	0.669 ± 0.045	0.784 ± 0.010	0.778 ± 0.012
	$\mathcal{O}(2^n)$	0.839 ± 0.009	0.848 ± 0.002	0.842 ± 0.002	0.781 ± 0.005	0.819 ± 0.004	0.827 ± 0.003	0.790 ± 0.004	0.821 ± 0.002
Qwen2.5-Omni-7B									
Linear	$\mathcal{O}(1)$	0.842 ± 0.005	0.829 ± 0.014	0.252 ± 0.106	0.248 ± 0.028	0.601 ± 0.024	0.440 ± 0.033	0.268 ± 0.023	0.497 ± 0.024
	$\mathcal{O}(1)$ **- Ours**	0.833 ± 0.004	0.837 ± 0.002	0.499 ± 0.034	0.473 ± 0.056	0.688 ± 0.008	0.572 ± 0.022	0.482 ± 0.058	0.626 ± 0.024
	$\mathcal{O}(n)$	0.837 ± 0.003	0.846 ± 0.005	0.839 ± 0.008	0.364 ± 0.020	0.711 ± 0.012	0.721 ± 0.015	0.393 ± 0.018	0.673 ± 0.010
	$\mathcal{O}(n+1)$	0.838 ± 0.003	0.845 ± 0.002	0.832 ± 0.007	0.759 ± 0.011	0.773 ± 0.007	0.795 ± 0.009	0.779 ± 0.004	0.803 ± 0.001
	$\mathcal{O}(2^n)$	0.842 ± 0.005	0.852 ± 0.003	0.841 ± 0.002	0.780 ± 0.004	0.820 ± 0.002	0.821 ± 0.006	0.789 ± 0.002	0.821 ± 0.001
MLP	$\mathcal{O}(1)$	0.842 ± 0.008	0.824 ± 0.021	0.183 ± 0.064	0.250 ± 0.049	0.498 ± 0.051	0.355 ± 0.063	0.260 ± 0.048	0.459 ± 0.035
	$\mathcal{O}(1)$ **- Ours**	0.844 ± 0.004	0.843 ± 0.004	0.391 ± 0.033	0.454 ± 0.064	0.649 ± 0.015	0.565 ± 0.024	0.468 ± 0.057	0.602 ± 0.025
	$\mathcal{O}(n)$	0.838 ± 0.007	0.842 ± 0.011	0.835 ± 0.003	0.331 ± 0.045	0.690 ± 0.040	0.675 ± 0.035	0.360 ± 0.047	0.653 ± 0.018
	$\mathcal{O}(n+1)$	0.837 ± 0.011	0.840 ± 0.004	0.831 ± 0.009	0.758 ± 0.029	0.773 ± 0.011	0.797 ± 0.009	0.779 ± 0.019	0.802 ± 0.008
	$\mathcal{O}(2^n)$	0.842 ± 0.008	0.852 ± 0.007	0.840 ± 0.007	0.773 ± 0.003	0.817 ± 0.007	0.822 ± 0.004	0.782 ± 0.007	0.819 ± 0.003
Qwen3-Omni-30B									
Linear	$\mathcal{O}(1)$	0.846 ± 0.003	0.729 ± 0.015	0.672 ± 0.060	0.727 ± 0.019	0.786 ± 0.016	0.702 ± 0.007	0.724 ± 0.014	0.741 ± 0.007
	$\mathcal{O}(1)$ **- Ours**	0.834 ± 0.002	0.793 ± 0.021	0.328 ± 0.024	0.368 ± 0.057	0.699 ± 0.006	0.291 ± 0.046	0.331 ± 0.050	0.520 ± 0.028
	$\mathcal{O}(n)$	0.843 ± 0.004	0.841 ± 0.007	0.843 ± 0.003	0.478 ± 0.066	0.752 ± 0.012	0.662 ± 0.067	0.455 ± 0.045	0.696 ± 0.023
	$\mathcal{O}(n+1)$	0.841 ± 0.005	0.840 ± 0.004	0.840 ± 0.002	0.782 ± 0.005	0.795 ± 0.006	0.790 ± 0.005	0.785 ± 0.003	0.811 ± 0.003
	$\mathcal{O}(2^n)$	0.846 ± 0.003	0.848 ± 0.004	0.844 ± 0.004	0.812 ± 0.004	0.825 ± 0.003	0.812 ± 0.003	0.806 ± 0.006	0.828 ± 0.003
MLP	$\mathcal{O}(1)$	0.843 ± 0.005	0.460 ± 0.304	0.412 ± 0.139	0.502 ± 0.286	0.711 ± 0.019	0.435 ± 0.306	0.449 ± 0.275	0.545 ± 0.178
	$\mathcal{O}(1)$ **- Ours**	0.840 ± 0.002	0.677 ± 0.123	0.171 ± 0.030	0.332 ± 0.124	0.657 ± 0.020	0.189 ± 0.057	0.288 ± 0.097	0.451 ± 0.055
	$\mathcal{O}(n)$	0.841 ± 0.006	0.843 ± 0.004	0.845 ± 0.005	0.433 ± 0.057	0.740 ± 0.013	0.552 ± 0.046	0.439 ± 0.037	0.671 ± 0.015
	$\mathcal{O}(n+1)$	0.842 ± 0.005	0.841 ± 0.004	0.844 ± 0.003	0.787 ± 0.016	0.763 ± 0.025	0.794 ± 0.002	0.797 ± 0.009	0.810 ± 0.007
	$\mathcal{O}(2^n)$	0.843 ± 0.005	0.846 ± 0.006	0.843 ± 0.005	0.808 ± 0.008	0.821 ± 0.012	0.810 ± 0.008	0.803 ± 0.006	0.825 ± 0.003
Phi-4-multimodal									
Linear	$\mathcal{O}(1)$	0.842 ± 0.005	0.654 ± 0.038	0.305 ± 0.120	0.056 ± 0.009	0.581 ± 0.073	0.042 ± 0.009	0.055 ± 0.013	0.362 ± 0.030
	$\mathcal{O}(1)$ **- Ours**	0.837 ± 0.003	0.712 ± 0.008	0.276 ± 0.020	0.186 ± 0.012	0.235 ± 0.014	0.180 ± 0.015	0.179 ± 0.011	0.372 ± 0.009
	$\mathcal{O}(n)$	0.843 ± 0.002	0.832 ± 0.005	0.842 ± 0.003	0.321 ± 0.038	0.577 ± 0.019	0.343 ± 0.046	0.334 ± 0.044	0.585 ± 0.019
	$\mathcal{O}(n+1)$	0.834 ± 0.003	0.828 ± 0.006	0.841 ± 0.003	0.750 ± 0.006	0.632 ± 0.015	0.786 ± 0.004	0.761 ± 0.003	0.776 ± 0.002
	$\mathcal{O}(2^n)$	0.842 ± 0.005	0.841 ± 0.003	0.843 ± 0.004	0.758 ± 0.006	0.800 ± 0.002	0.799 ± 0.008	0.764 ± 0.008	0.807 ± 0.002
MLP	$\mathcal{O}(1)$	0.840 ± 0.005	0.625 ± 0.137	0.376 ± 0.158	0.035 ± 0.041	0.556 ± 0.085	0.027 ± 0.033	0.032 ± 0.039	0.356 ± 0.035
	$\mathcal{O}(1)$ **- Ours**	0.837 ± 0.003	0.606 ± 0.010	0.142 ± 0.044	0.150 ± 0.036	0.199 ± 0.025	0.130 ± 0.029	0.142 ± 0.035	0.315 ± 0.013
	$\mathcal{O}(n)$	0.835 ± 0.003	0.837 ± 0.007	0.846 ± 0.002	0.261 ± 0.074	0.557 ± 0.038	0.258 ± 0.064	0.266 ± 0.078	0.551 ± 0.036
	$\mathcal{O}(n+1)$	0.841 ± 0.006	0.833 ± 0.008	0.843 ± 0.010	0.728 ± 0.027	0.583 ± 0.029	0.791 ± 0.004	0.759 ± 0.015	0.768 ± 0.009
	$\mathcal{O}(2^n)$	0.840 ± 0.005	0.842 ± 0.007	0.838 ± 0.003	0.761 ± 0.007	0.799 ± 0.006	0.797 ± 0.009	0.764 ± 0.006	0.806 ± 0.002
Interactive-Omni-8B									
Linear	$\mathcal{O}(1)$	0.844 ± 0.004	0.793 ± 0.028	0.330 ± 0.152	0.410 ± 0.077	0.526 ± 0.076	0.279 ± 0.098	0.363 ± 0.086	0.506 ± 0.070
	$\mathcal{O}(1)$ **- Ours**	0.830 ± 0.002	0.829 ± 0.002	0.545 ± 0.020	0.682 ± 0.012	0.651 ± 0.012	0.674 ± 0.007	0.653 ± 0.006	0.695 ± 0.008
	$\mathcal{O}(n)$	0.841 ± 0.003	0.840 ± 0.005	0.834 ± 0.005	0.459 ± 0.035	0.708 ± 0.023	0.705 ± 0.037	0.590 ± 0.050	0.711 ± 0.020
	$\mathcal{O}(n+1)$	0.842 ± 0.002	0.838 ± 0.005	0.831 ± 0.006	0.693 ± 0.037	0.782 ± 0.009	0.803 ± 0.006	0.780 ± 0.005	0.796 ± 0.004
	$\mathcal{O}(2^n)$	0.844 ± 0.004	0.845 ± 0.004	0.845 ± 0.003	0.801 ± 0.005	0.816 ± 0.003	0.822 ± 0.004	0.788 ± 0.004	0.823 ± 0.000
MLP	$\mathcal{O}(1)$	0.839 ± 0.006	0.741 ± 0.059	0.191 ± 0.069	0.327 ± 0.081	0.442 ± 0.065	0.208 ± 0.098	0.297 ± 0.102	0.435 ± 0.064
	$\mathcal{O}(1)$ **- Ours**	0.837 ± 0.001	0.828 ± 0.001	0.462 ± 0.035	0.688 ± 0.038	0.650 ± 0.018	0.650 ± 0.024	0.680 ± 0.016	0.685 ± 0.018
	$\mathcal{O}(n)$	0.840 ± 0.008	0.844 ± 0.006	0.836 ± 0.010	0.453 ± 0.041	0.734 ± 0.017	0.721 ± 0.068	0.609 ± 0.051	0.720 ± 0.024
	$\mathcal{O}(n+1)$	0.838 ± 0.005	0.835 ± 0.005	0.834 ± 0.007	0.696 ± 0.045	0.765 ± 0.014	0.805 ± 0.006	0.787 ± 0.009	0.794 ± 0.007
	$\mathcal{O}(2^n)$	0.839 ± 0.006	0.846 ± 0.007	0.841 ± 0.010	0.797 ± 0.007	0.819 ± 0.003	0.821 ± 0.004	0.795 ± 0.003	0.823 ± 0.002